

隠れマルコフモデルによる歴史テキストの人物移動のモデル化

Movement Modeling for History Documents with Hidden Markov Models

水谷 陽太 *1

Yota Mizutani

鶴岡 慶雅 *2

Yoshimasa Tsuruoka

*1 東京大学大学院工学系研究科

Graduate School of Engineering, The University of Tokyo

*2 東京大学大学院情報理工学系研究科

Graduate School of Information Science and Technology, The University of Tokyo

It is difficult for computers to understand the “meaning” of natural language sentences. To tackle this problem, some existing methods use predicate logic. However, they cannot deal with quantitative data such as geographical distance which is important for understanding historical events. We introduce a new method to construct a simulatable world model from documents. Simulations with this model will help computers to understand the contexts, guess unwritten information, and realize some rules. We experiment with some documents about the Sengoku period in Japanese Wikipedia, and construct a hidden Markov model about people’s movement.

1. はじめに

品詞タグ付けや構文解析と言った自然言語処理の基盤的技術は、近年の機械学習技術の発展に伴い大幅に精度が向上し、分野適応や学習データの作成コストの問題は依然として残るものの、テキスト処理のための実用的な技術として確立しつつある。しかし、自然言語の「意味」の理解を必要とする技術、すなわち（広い意味での）質疑応答や、（知的なエージェントを実現するための）自然言語理解などに関しては、現在の自然言語処理技術によってコンピュータで実現可能なことは極めて限られている。

近年、「ロボットは東大に入れるか？」プロジェクト [新井 12] を始めとして、計算機による自然言語理解の実現に対する注目は高まっている。しかし、現在の言語処理システムでは、例えば、「1467 年の応仁の乱で京都の大半が消失した」というテキストから、「15 世紀末に京都の人口が減少した」ということを推論することはほぼ不可能である。この種の世界知識を必要とする推論に関する問題は、人工知能研究の黎明期から研究されており、古くは Winograd による SHRDLU [Winograd 72] や、最近では、カーネギーメロン大学の Read the Web プロジェクト [Carlson 10] などによって実現可能なアプローチが模索されている。しかし、計算機による自然言語理解という問題に対する既存のアプローチのほとんどは述語論理をベースとした知識表現や推論機構を用いており、「世界」に関する定量的な情報や、時間発展の現象を扱うことが極めて難しいという問題がある。このことは、先に例としてあげたような歴史的事象の理解を目的とする場合には致命的な欠点であり、どれだけ大量に知識を積み重ねても、いわゆる「風が吹けば桶屋が儲かる」といったような、論理的にはつながっていても現実的には起こり得ない推論をしてしまう問題を避けることができない。また、歴史テキストを解釈するためには、歴史イベント同士の地理的な（遠近の）関係などもしばしば考慮する必要があるが、この種の定量的な情報はそもそも述語論理で表現することが適さない。

そこで本研究では、世界史や日本史といった、歴史的事象に
連絡先: 水谷陽太, 東京大学大学院工学系研究科電気系工学専攻, mizutani@logos.t.u-tokyo.ac.jp

関するテキストを理解・解釈することのできる計算機システムを実現するため、様々な歴史イベントを表現可能な、自律的に時間発展する世界モデルを用いることを提案する。テキストに記述された歴史イベントを読み取り、記述された内容に即した世界モデルを学習により獲得する。世界モデルは、歴史的事象を表現可能であるだけでなく、その時間発展が、史実、すなわちテキストの内容に合致することが求められる。学習により得られた世界モデルは、文脈の判断 [村上 16] や、言外の情報の推測、より本質的な原理の理解など、「意味」を理解した言語処理の助けになることが期待される。

将来的には人物、地理、時間等の多数のパラメータに依存して発生する様々な種類のイベントを扱うことが目標だが、本稿ではその前段階として、人物の移動のみに注目して実験を行った。EM アルゴリズムの一種である Baum-Welch アルゴリズム [Baum 70] を用いて特定の地点間の遷移確率を推定し、隠れマルコフモデルを構築する実験を行った。その結果、データ不足に起因する偏りは確認されたものの、テキストから読み取ったイベントを表現するモデルを獲得することに成功した。

2. 提案手法

本手法では人物の移動モデルを、移動先がその人物の現在位置のみに依存する単純マルコフ過程であると仮定する。時刻 t で特定の位置 p_i に存在するとき、 $t+1$ で p_j に移動する確率を表す遷移行列 A_{ij} を求めることが目的となる。ある人物の様子が記述されたテキストを解釈することで、その人物が時刻 t において位置 p に存在したという記述 (t, p) を得る事ができる。ただし、全ての t について (t, p) がテキストから得られることは稀であり、記述が存在しない t については、位置の推定が必要となる。この問題に対し提案手法では、記述の有無を含めた移動モデルを隠れマルコフ過程で表現し、Baum-Welch アルゴリズム [Baum 70] を用いて遷移行列の推定を行う。より具体的には、時刻 t において人物が特定の位置 p_i に存在しているとき、観測として「 (t, p_i) の記述が得られる」場合と、「 t における記述が得られない」場合の二通りが存在すると考える。一般的な隠れマルコフモデルでは、この他に i と異なる j についても (t, p_j) が得られる可能性を考慮し、またこれらの観

測を得る確率は現在位置 p_i によって異なることを仮定するが、本稿では誤った記述 (t, p_j) は存在しないことを仮定し、また、記述が行われる確率は現在位置によらず一定であるとした。

2.1 アルゴリズム

N 人の人物の各 (t, p) のリストを入力とし、遷移行列 A_{ij} が出力となる。通常の Baum-Welch アルゴリズムとは異なり、観測シンボル確率分布は固定であることに注意する。

- 遷移行列 A_{ij} をランダムに初期化する。
- 特定の人物について、 (t, p) のリストから観測シンボル列を生成する。
- 現在のパラメータ A_{ij} を用いて、通常の Baum-Welch アルゴリズムと同様に期待値計算を行う。
- 他の人物においても同様に観測シンボル列の生成及び期待値計算を行う。
- N 人全ての処理が終わったら、Baum-Welch アルゴリズムに従いパラメータ A_{ij} の更新を行う。

以上の操作を 1 epoch とし、複数回繰り返す。

3. 実験

日本の戦国時代の人物に関して、日本語 Wikipedia^{*1} の記述から遷移行列を求める実験を行った。

3.1 データの作成

日本語 Wikipedia の「戦国時代の人物一覧 (日本)」ページに挙げられた人物について、リンク先の各人物について記述されたページ内の記述から、データ生成を行った。時間の単位は一ヶ月とした。文単位でパターン検索を行い、一つの文内に時刻のデータと地名のデータが両方存在する場合、それらの組をデータに加える。時刻は一ヶ月単位で保存し、もし年のみの記述が存在し、月が不明の場合は暫定的に 6 月であるとした。地名に関しては、日本語 Wikipedia の「令制国」ページ及び「日本の城一覧」ページから地名の一覧を作成した。ただし、データ取得後、「令制国」ページの「変遷」の章にあるデータから、取得した地名を戦国時代に使われていた国名へ変換した。同様に、「日本の城一覧」ページ内のデータから、城の名前を所在地のデータへと変換した。なお、本実験では人物の移動に着目するため、時刻と地名のデータ対が 2 組以上存在した人物のデータのみを用いた。取得されたデータ数を表 1 に示した。地名は実際に使用されるデータ中に存在するものの数である。

表 1: 実験に用いたデータ数

人物	地名	データ数
1189	68	4880

例えば「織田信長」のページでは、「元亀 2 年 (1571 年) 2 月、信長は浅井長政の配下の磯野員昌を降し、佐和山城を得た。」という文から、(1571 年 2 月, 佐和山城) というデータを得ることができる。この際、佐和山城は所在地である「近江国」へと変換され、(1571 年 2 月, 近江国) というデータとして扱われる。表 2 の太字部分がこのデータにあたる。

3.2 結果

Baum-Welch アルゴリズムを 10000 epoch 実行し、遷移行列の推定を行った。推定された遷移行列を用いたシミュレーションの例として、織田信長のデータの欠損部分を最尤推定した結果の一部を、表 2 に示す。

3.3 考察

概ね隣接する国間を移動する様子が推定できている。例えば 1571 年 6 月の三河国から 1572 年 6 月の摂津国へと移動をする際は、地理的にそれらの二国間に存在する近江国、山城国を経由する推定を行っている。同様に、1577 年 2 月の紀伊の国から 1577 年 12 月の美作国へと移動する際も、大和国、摂津国、播磨国を経由する推定となった。

反面、1572 年 6 月の摂津国から 1573 年 2 月の陸奥国への移動の際は、途中経路が全く推定できていない。今回の手法では離れた場所間の移動経路を推定するのは難しいことが読み取れる。ただし、このデータは本来三河国の野田城を指すテキストを処理する際に、陸奥国にある同名の野田城であると誤認したものである。今回の実験におけるデータ生成は単純なパターンマッチにより行っているため、このように不正確なノイズが入りやすくなっている。

4. おわりに

本稿では、歴史上のイベントについて記述されたテキストから、シミュレーション可能なモデルを構築し、テキストの内容を理解、解釈する手法を提案し、実際に人物の移動というイベントに関して、Baum-Welch アルゴリズムを用いた遷移確率の推定を用いてモデル化を行った。

今後の課題としては、データ数の増加、データの正確性の向上が挙げられる。今回は単純なパターン検索によるデータ作成を行ったが、文構造の解析を組み合わせるなどの手法によって、より正確なデータを作成できる。また、文脈を考慮することにより、文をまたいだデータを追加することが可能となる。完全な構文解析や文脈判断は困難だが、本稿の手法により得られた世界モデルを用いることで、それらの精度向上が見込める。手法を組み合わせることで相互に精度を向上させることが期待できる。

また、本稿では人物の移動のみに注目したシミュレーションを行ったが、季節や地理、人物の家系、地位、など様々なパラメータの組み込みを検討し、扱えるイベントの種類を増やすことで、より高度かつ広範囲な応用が期待される。

参考文献

- [Baum 70] Baum, L. E., Petrie, T., Soules, G., and Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The annals of mathematical statistics*, Vol. 41, No. 1, pp. 164–171 (1970)
- [Carlson 10] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M.: Toward an architecture for never-ending language learning., in *AAAI*, Vol. 5, p. 3Atlanta (2010)
- [Winograd 72] Winograd, T.: Understanding natural language, *Cognitive psychology*, Vol. 3, No. 1, pp. 1–191 (1972)

*1 <https://ja.wikipedia.org/wiki/>

表 2: 織田信長の移動経路推定

日時	既知	推定	推定確率	日時	既知	推定	推定確率
1570 年 9 月		越前国	0.504074	1574 年 11 月		越前国	0.988014
1570 年 10 月		近江国	0.525936	1574 年 12 月		越前国	0.993344
1570 年 11 月		近江国	0.629295	1575 年 1 月	越前国		1
1570 年 12 月		近江国	0.741078	1575 年 2 月		大和国	0.989675
1571 年 1 月		近江国	0.862448	1575 年 3 月	摂津国		1
1571 年 2 月	近江国		1	1575 年 4 月	三河国		1
1571 年 3 月		近江国	0.682636	1575 年 5 月		三河国	0.898059
1571 年 4 月		三河国	0.498432	1575 年 6 月		三河国	0.795449
1571 年 5 月		三河国	0.740088	1575 年 7 月		三河国	0.691781
1571 年 6 月	三河国		1	1575 年 8 月		三河国	0.586337
1571 年 7 月		三河国	0.838587	1575 年 9 月		三河国	0.478777
1571 年 8 月		三河国	0.687775	1575 年 10 月		近江国	0.551551
1571 年 9 月		三河国	0.548078	1575 年 11 月		近江国	0.681435
1571 年 10 月		三河国	0.420206	1575 年 12 月		近江国	0.828343
1571 年 11 月		三河国	0.305626	1576 年 1 月	近江国		1
1571 年 12 月		三河国	0.206029	1576 年 2 月		近江国	0.70717
1572 年 1 月		近江国	0.187306	1576 年 3 月		近江国	0.433539
1572 年 2 月		近江国	0.162148	1576 年 4 月		山城国	0.226821
1572 年 3 月		山城国	0.20815	1576 年 5 月		摂津国	0.382394
1572 年 4 月		摂津国	0.2867	1576 年 6 月	摂津国		1
1572 年 5 月		摂津国	0.418255	1576 年 7 月		摂津国	0.491541
1572 年 6 月	摂津国		1	1576 年 8 月		摂津国	0.406495
1572 年 7 月		摂津国	0.429042	1576 年 9 月		摂津国	0.363946
1572 年 8 月		摂津国	0.303165	1576 年 10 月		摂津国	0.365772
1572 年 9 月		陸奥国	0.375088	1576 年 11 月		摂津国	0.365958
1572 年 10 月		陸奥国	0.498288	1576 年 12 月		山城国	0.454871
1572 年 11 月		陸奥国	0.613336	1577 年 1 月		山城国	0.936523
1572 年 12 月		陸奥国	0.728341	1577 年 2 月	紀伊国		1
1573 年 1 月		陸奥国	0.84351	1577 年 3 月		大和国	0.572397
1573 年 2 月	陸奥国		1	1577 年 4 月		摂津国	0.428538
1573 年 3 月		陸奥国	0.849924	1577 年 5 月		摂津国	0.288138
1573 年 4 月		陸奥国	0.736186	1577 年 6 月		播磨国	0.293367
1573 年 5 月		陸奥国	0.599787	1577 年 7 月		播磨国	0.382155
1573 年 6 月		山城国	0.418539	1577 年 8 月		播磨国	0.432742
1573 年 7 月	山城国		1	1577 年 9 月		播磨国	0.519875
1573 年 8 月	山城国		1	1577 年 10 月		播磨国	0.510755
1573 年 9 月		紀伊国	0.512643	1577 年 11 月		播磨国	0.831622
1573 年 10 月		近江国	0.590877	1577 年 12 月	美作国		1
1573 年 11 月	近江国		1	1578 年 1 月		播磨国	0.999488
1573 年 12 月		越前国	0.504339	1578 年 2 月		播磨国	0.596002
1574 年 1 月	越前国		1	1578 年 3 月	播磨国		1
1574 年 2 月		越前国	0.993344	1578 年 4 月		播磨国	0.590693
1574 年 3 月		越前国	0.988014	1578 年 5 月		播磨国	0.577634
1574 年 4 月		越前国	0.983885	1578 年 6 月		播磨国	0.465685
1574 年 5 月		越前国	0.980956	1578 年 7 月		播磨国	0.389094
1574 年 6 月		越前国	0.979208	1578 年 8 月		播磨国	0.312321
1574 年 7 月		越前国	0.978625	1578 年 9 月		播磨国	0.245015
1574 年 8 月		越前国	0.979208	1578 年 10 月		摂津国	0.22699
1574 年 9 月		越前国	0.980956	1578 年 11 月		摂津国	0.223781
1574 年 10 月		越前国	0.983885	1578 年 12 月		摂津国	0.212737

- [新井 12] 新井 紀子, 松崎 拓也 他: ロボットは東大に入れるか?, 人工知能学会誌, Vol. 27, No. 5, pp. 463–469 (2012)
- [村上 16] 村上 優樹, 鶴岡 慶雅 他: 現実世界の時間・空間制約を用いた共参照解析の精度向上, 言語処理学会第 22 回年次大会, pp. 250–253 (2016)