ベイジアン隠れマルコフモデルとWikipediaテキストを用いた歴史人物移動モデルの推定

古川 好　　鶴岡 慶雅
東京大学 工学部電子情報工学科
{furukawa,tsuruoka}@logos.t.u-tokyo.ac.jp

1 はじめに

自然言語処理は様々な分野に分かれているが、その中の一つには自然言語理解がある。自然言語理解とは人間が言語を理解するのと同じように、コンピュータもテキストの表層的な情報だけでなく、事前知識を必要とする深層的な情報まで読み取ることができるようになることを目指す試みである。

自然言語理解に関して、水谷ら [1] や村上ら [2] は、テキストから学習を行い現実世界の地理的・時間的なモデルを作成する手法や、質疑応答や文章の正誤判定に現実世界の地理的・時間的な関係を考慮する手法を提案している。例えば、地名の空間における位置を表現するモデルを作ることができたとする。その時、「私は 1 時に北海道におり、1 時 1 分に沖縄にいた。」という文章の正誤判断を行う際、北海道と沖縄の地理的な関係を知らない場合は判断することができないが、前述の空間的・時間的モデルを用いることによって地理的な関係を考慮できれば容易に解くことが可能である。また、このようなモデルを作成することができれば、過去の人間が移動する際にどのような経路を辿ったのかを考証する際などにも使用でき、様々な応用が可能だと考えられる。

本研究では、日本語版 Wikipedia のテキストから歴史的な人物がどの時点でどの場所にいたかというイベントを抽出し、そのデータからイベントとイベントの間の欠落において、人物がどのように移動を行っていたのかを推測するような人物移動モデルを作成するという水谷ら [1] の研究を元とし、より精度の高いモデルの作成を目指した。水谷らの研究において問題とされていたのは、人物移動モデルを学習する際のデータ不足、イベントの抽出法に起因するデータの不正確性の 2 つであり、これらの問題点に対してモデルの変更及び、文脈情報を利用したイベント抽出を行うことによって解決を試みた。結果として、データの数及びデータの正確性を向上させることに成功した。

2 関連研究

ここでは、本研究の下地である水谷らの作成した歴史人物移動モデル [1] について説明する。日本語版 Wikipedia のテキストから、人物がどのように移動したか、つまりある時刻 t においてある位置 p にいたという情報の組 (t, p) を、個々の人物毎に抽出してデータセットを作成する。具体的には以下のような手順で情報の抽出とイベント (t, p) の作成を行った。

・日本語版 Wikipedia の全記事を取得し、人物の移動を令制国単位でなされるものとするために、日本語 Wikipedia の「令制国」及び「日本の城一覧」のページより現在の都道府県名や城名、戦国時代以外の時代での令制国名と戦国時代の令制国名を対応させた辞書を作成する。

・「日本の元号」のページから日本の元号と西暦を対応させた辞書を作成する。

・「戦国時代の人物一覧 (日本)」のページに掲載されている人物が戦国時代の人物であるとし、その人物達をまとめたリストを作成する。

・リストにまとめられた全人物に対してその人物のページから (t, p) の抽出を行う。パターンマッチを行い、作成した地名と年号に関する辞書を用いて、一つの文内に時刻のデータと地名のデータが共に存在するかを調べ、存在する場合はそれらの組を抽出してデータとする。この時、時刻は一ヶ月単位とし、年はわかるが月が不明な場合は 6 月と仮定する。例えば、「上杉謙信」のページにおいて、「天正 5 年（1577 年）12 月 18 日、謙信は春日山城に帰還し、12 月 23 日には次なる遠征に向けての大動員令を発した。」という一文があった場合、辞書によって天正 5 年 12 月が時刻 $(1577 \times 12 + 12) = 18936$ に、春日山城が越後国に変換され、(18936, 越後国) という組として抽出される。

こうして得られたイベント (t, p) を隠れマルコフモデルの観測とし、得られなかった時刻のイベントについては隠れマルコフモデルの状態が「記述なし」を出力したと考えることによって、人物移動モデルの学習を行った。隠れマルコフモデルの学習には最尤推定法の一種である Baum-Welch アルゴリズムを用いる。

3 提案手法

先行研究において問題点とされていたのは、学習に使用できるデータが少ないことと、データの正確性が低いことである。そこで、解決のために以下の手法を提案する。

3.1 データ数の増加

イベント抽出の対象となる人物を、戦国時代の人物だけではなくほぼ同時代の安土桃山時代の人物も含めることで、抽出できるイベントを増加させる。あまりにも時代が離れすぎていてしまうと、交通手段や地名の変化により正しい推定が行うことができなくなってしまうが、近い時代であればその影響も少ないと考えられる。更に、パターンマッチに反応する地名を増加させるために、城名と令制国名だけでなく寺社名や戦争名も地名として扱うことができるようにした。

3.2 データの正確性の向上

先行研究におけるイベント抽出では、テキストの文脈や文構造が考慮されていなかった。そこで、本研究では以下に示す二通りの方法を用いて適切なイベントの抽出を試みた。

一つめは、文章解析ソフト KNP [3] を用いることによって文章の主体や文脈情報を解析し、その情報を用いてルールベースの制約条件を作成することで誤ったデータを抽出する可能性を減らすというものである。例えば、文章に存在する主体が、その文章の記載されている記事の主役と一致しない場合は、その文章は記事の人物について書かれたものではないと判断し、イベント抽出の対象としない、などのような制約である。

二つめは、イベント抽出を文章の分類タスクとして捉える手法である。イベントが適切なものであるかどうかを見極める際に重要となるのは、多くの場合はその文章の主体と文章中に登場する地点である。そこで、ここではある文章とその文章の存在する記事の主役から、その文章の主体と記事の主役が一致するかどうかを判別する二値分類問題と、登場する地名にその文の主体が存在していると文章から読み取ることができるかどうかを判別する二値分類問題を解き、その両方が条件を満たした場合のみ適切なイベントとして抽出をすることにする。

3.3 モデルの学習方法の変更

隠れマルコフモデルの学習に、Baum-Welch アルゴリズムではなくベイズ推論を用いることを提案する。Baum-Welch アルゴリズムを用いた隠れマルコフモデルの場合、訓練データに存在していないような遷移を起こす確率はどれだけ学習を行っても 0 のままであり、訓練データにはないが適切だと考えられるような遷移を表現することはできない。しかし、ベイズ推定では初期確率行列などのパラメータそのものを学習するのではなく、パラメータを生成する確率分布を学習するため、遷移確率行列は最尤推定の場合と比べてスムージングされるので確率が 0 となることはない。故に、本研究の場合のように訓練データ数が少ない場合にはベイズ推定により学習をした場合の方が有効であると考えられる。Beal らの研究 [4] によると、式 (1) による更新を繰り返し行うことにより、EM アルゴリズムにおける反復計算のようにベイジアン HMM のパラメータの更新を実行できることが知られている。θ が遷移確率行列の更新式、ϕ が出力確率行列の更新式である。ただし、$n_{S',S}$ は状態 S から $S^{'}$ に遷移した回数、n_S は状態 S であった回数、$n^{'}_{o,S}$ は状態 S で o を出力した回数、m、$m^{'}$ は状態及び出力の数、$\Psi(x)$ はディガンマ関数を表している。

$$\theta = \frac{\exp(\Psi(E[n_{S',S}] + \alpha))}{\exp(\Psi(E[n_S] + m\alpha))} \tag{1}$$

$$\phi = \frac{\exp(\Psi(E[n^{'}_{o,S}] + \alpha^{'}))}{\exp(\Psi(E[n_S] + m\alpha^{'}))} \tag{2}$$

4 実験

4.1 データの作成

イベントの抽出には、日本語版 Wikipedia の「戦国時代の人物一覧」及び「安土桃山時代の人物一覧」のページに掲載されている人名についての記事を使用し

表 1: 得られたデータ数

	人物数	地名数	データ数
既存手法	1263	69	5121
提案手法 (KNP)	838	70	3330
提案手法 (BERT)	1090	70	4567

た。3 節で述べられている記事の主役とは Wikipedia 記事のタイトルを表している。更に、「令制国」「日本の城一覧」「日本の寺院一覧」「日本の合戦一覧」のページより地名の一覧を作成し、「令制国」ページ内に存在する戦国時代に使われていた旧国名と対応付ける辞書を作成した。

KNP を用いて文脈解析を行い、その情報を用いてルールベースでイベント抽出を行う場合は、テキストの分かち書きを行う際に予め作成した人名や地名の一覧を事前知識として使用することで、後の文脈解析を行いやすくした。KNP によって得られた情報を用いて、先述したようなルールベースの制約条件を課すことによって、イベント抽出を行った。

イベント抽出を二値分類問題と考える場合は、まず二値分類に必要となるデータセットを作成した。文章と主体、もしくは文章と地名の組み合わせが正しいものであるかは人手でアノテーションを行い、それぞれ 1000、1200 個ずつデータを作成した。このデータセットを用いて BERT [5] と呼ばれる言語表現モデルの転移学習を行い、残りのデータに対しラベル付けを行った。

4.2 結果

表 1 が得られたデータ数である。既存手法は水谷らによる研究 [1] の場合を表している。

また、得られたデータを使用し、既存手法の場合は最尤推定による学習で隠れマルコフモデルを 10000 エポック学習し、提案手法の場合はベイズ推定による学習で同様に 10000 エポック学習した。得られた結果の一例として、上杉謙信の 1573 年の移動推定を行ったものを以下の表 2 に示す。

4.3 考察

表 1 から分かるように、結果としてデータの総数自体は既存手法の場合と比べて提案手法の方が少なくなってしまっている。これは、既存手法におけるパターンマッチでは誤ったデータがかなり多く、使用するデータを増やしたとしても結果的には提案手法により不適切なものだと判断されてしまうものが多いからだと考えられる。表 2 によると、3 月の部分では既存手法のみが越中国がイベントとして得られたとしているが、これは「翌天正元年（1573 年）3 月、信玄の画策により再起した越中一向一揆が再度富山城を奪った。」という文章を誤って抽出してしまったものである。一方で提案手法では陸奥国を抽出しており、これは「3 月には膳城・女淵城・深沢城・山上城・御覧田城を立て続けに攻め落とし戦果をあげた。しかし成繁の居城である要害堅固な金山城を陥落させるに至らず。」という文章を正しく抽出したものであり、文脈の理解が正しいイベントの抽出に役立っていることが分かる。その一方で、BERT の場合は 12 月に大和国がイベントとして得られたとしているが、これは「12 月 19 日、剃髪して法印大和尚に任ぜられる。」という文章の「大和」の部分を誤って抽出したものであり、完全に正しいイベントのみを抽出できているとは言えない。これは、BERT の学習に使用したデータセットの数が少なく不十分であることや、BERT の事前学習モデルとして使用している多言語モデルにおいて日本語の扱いが雑であり、正しい事前学習が行えていない可能性があることなどが大きな原因として考えられる。

5 おわりに

本実験においては歴史上の人物についてのテキストから、人物がどの時刻 t でどの地点 p にいたのかを自然言語処理的な手法を用いて抽出し、そうして得られたデータを隠れマルコフモデルの観測と考えてベイズ推定による学習を行うことで、人物移動モデルを作成した。

今後の課題としては、まず先行研究でも指摘されていたようにデータ不足が挙げられる。本研究に置いては参照する時代の拡大や、イベント抽出の際に参照する地名を増やすことによってデータの増量を試みたが、適切でないイベントを KNP や BERT を用いて除外した結果、正確なデータの割合は向上したがデータ自体の総数は増加しないという結果になってしまった。次に挙げられるのは、人物移動モデルに更なる情報を取り入れるというものである。地点間に存在する山などの空間的な情報を事前知識として活用することで、より精度の高い人物移動を推測することができるようになると考えられる。

表 2: 人物移動モデルの推定例

	既存手法			提案手法 (KNP)			提案手法 (BERT)		
日時	既知位置	推定位置	推定確率	既知位置	推定位置	推定確率	既知位置	推定位置	推定確率
1 月		信濃国	0.3478		越中国	0.4333		摂津国	0.1524
2 月		信濃国	0.4993		越中国	0.2870		摂津国	0.1387
3 月	越中国		1	陸奥国		1	陸奥国		1
4 月		越後国	0.4371		陸奥国	0.4252	武蔵国		1
5 月		信濃国	0.3735	越後国		1	越後国		1
6 月		信濃国	0.3232		越後国	0.5337		越後国	0.3766
7 月		越後国	0.5569		越後国	0.3014		出羽国	0.2935
8 月	上野国		1	越中国		1	越中国		1
9 月		相模国	0.5171		能登国	0.5605		出羽国	0.9748
10 月		上野国	0.2641		越中国	0.6284	武蔵国		1
11 月		相模国	0.2714		能登国	0.3523		相模国	0.5119
12 月		信濃国	0.2013		越中国	0.4333	大和国		1

参考文献

[1] 水谷陽太, 鶴岡慶雅. 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. 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 4Pin1–17, 2018.

[2] 村上優樹, 鶴岡慶雅. 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 発表論文集, 2016.

[3] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. *Journal of Natural Language Processing*, Vol. 14, No. 4, pp. 67–81, 2007.

[4] Matthew J. Beal. Variational algorithms for approximate bayesian inference. Technical report, 2003.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv: 1810.04085*, 2018.