歴史上の人物志に対する属性・属性値の推定

本澤 拓[1]　森 信介[1,2]　関野 樹[3]

概要：人文学の資料について，情報を資料横断的に共通したデータ構造でまとめることは人文学の研究を円滑に進める上で重要である．本稿では人物の説明文を対象とし，より多くの内容を被覆するような属性と属性値の策定を行う．そして機械学習を用いて，それらの属性および属性値を抽出する自動抽出器の構築を試みる．

1. はじめに

5W1H といわれるように，誰（Who）は物事を語る上で必須の要素である．そして，その「誰」を識別するのに広く用いられるのが人名である．現代の人物であれば，SNS をはじめとしするインターネット上のサービスで人名を検索し，その人物に関する情報を集めることが可能である．しかしながら，歴史上の人物となると，人名およびその人物の関連情報を導き出すことは容易ではない．歴史人物は，人名自体が現代よりも複雑であり，姓名以外にも字，号などが人物の識別に用いられる．特定の歴史人物に関する情報を得ようとすれば，これらを史資料から抽出し，それぞれの人物と関連付けてゆく作業が必要である．このため，歴史人物の人名を収集し，そのデータを集約する取り組みが始まっている [1]．従来，これらの歴史人物に関する情報を抽出・集約する作業は人手で進められており，史資料を読み解く知識および膨大な時間と労力が作業の進捗の妨げとなってきた．そこで，本研究では，自然言語処理技術を歴史人物の情報の抽出に活用することを試みる．人物解説に関する史資料を題材とし，これに対するアノテーションとそのアノテーションデータを用いた推定を行う．これらに併せて，歴史人物に関する情報を集約するためのデータ構造についても検討を加える．さらに，本研究では，人物の関連情報として，時空間情報に注目する．人物の生没年，住居などの情報を人物開設から自動的に抽出する技術を構築し，人物を関連付ける．これにより，人物を時空間軸上に位置付けることが可能となり，当時の人間関係の傍証としたり，同名の複数の人物を弁別したりすることが可能となる．また，人物を時空間情報と関連付けることは，同様に時間や空間に関連付けることができる団体，建物，文書，事件などさまざまな事象と人物の関係を知るための手がかりともなり得る．

表 1　人物情報テキストと抽出対象の文字列の例．人物情報テキストの太字は抽出対象となる文字列を表す．

属性	内容
短冊番号	23
人物名	熊谷尚之
人物情報テキスト	熊谷尚之 (**寛政 11 年**) **漢学者**。字**履善**、号**箕山**、通称**平一郎**。**京都の**人、**新町二条下ル町**に住み**井上金蛾**に学んだ。かねて**詩**を作り**和歌**を善くした。**寛政十一年十月十日**没、年**七十一**。
字	履善
号	箕山
別名	平一郎
没年	寛政 11 年，寛政十一年十月十日
享年	七十一
出身地	京都
住所	新町二条下ル町
師匠	井上金蛾
職業	漢学者
得意分野	詩，和歌

2. 課題

本論文では，人物についての説明文に対して，その属性を策定するとともに，各説明文から属性値を自動的に収集するための手法を提案する．人物についての説明文としては，インターネットを通じて公開されている「平安人物志」と「平安人物志短冊帖」を題材とする．「平安人物志」は近世京都の文化人について記述した資料である．「平安人物志短冊帖」には，「平安人物志」に掲載された諸家の短冊に関する情報とともに，一部の諸家の人物情報が付されて

[1] 株式会社 Linfer
[2] 京都大学 学術情報メディアセンター
[3] 国際日本文化研究センター

いる．

本論文での課題は，

(1) 人物の説明文に対して，記述されている内容の多くを被覆する属性と属性値の定義を策定すること

(2) その定義に基づいて書く人物の説明文から属性・属性値を自動的に抽出する方法を提案すること

である．

表 1 の例では人物「熊谷尚之」の説明文と属性を示す．人物情報テキストの太字の文字列は抽出対象の文字列である．

3. 属性および属性値の定義

可能な限り多くの情報を自動抽出することを目指して属性および属性値を定義する．このために，まず芳賀人名に対して人出で策定された属性である別名，親，仕えた人，死没年月日，死没時齢を抽出した [1]．次にその結果を目視し，芳賀人名で策定された属性では被覆されない属性を列挙して，最終的に 21 種類の属性を決定とした．各属性に対して，属性値（文字列）の範囲を定義した．各人物に対してただ 1 つに決まるであろう属性についても，該当する文字列が複数あれば全て抽出対象とする．例えば表 1 では，没年に相当する「寛政 11 年」と「寛政十一年十月十日」という 2 つの文字列が説明文中に存在しているが，どちらも抽出対象とする．以下では，21 の属性を 5 つに分類して順に説明する．

3.1 名前

姓，名，字，号，その他別名の 5 属性を抽出した．姓，名，字，号はそれ以外の別名と比べて数が多いので，属性を分けた．諡や通称などは属性を分けるほど十分な数がなかったので，別名に分類した．近世の文化人は多くの別名を持っており，資料によっては同一人物が別の名前で登場している可能性もあり，別の資料との連携を図る上で別名は重要となる．

3.2 時間情報

生まれた年，没年，享年の 3 属性を抽出対象とした．生まれた年は他の 2 つの属性に比べて出現回数が少ないが，没年と享年が抽出できれば特定が可能である．生没年は人物の活動時期を特定する上で重要であり，住所などの詳細な空間情報と合わせれば人物の詳細な活動範囲を特定できる．

3.3 空間情報

出身地，住所，墓地の 3 属性を抽出対象とした．出身地は「京都」や「江戸」などのように広い範囲で記述されており，墓地は人物の活動場所とは関係がないので，人物の活動していた位置を正確に特定するのにはそれほど寄与しない．一方住所は「京都」や「江戸」など広い範囲で記述されているものもあれば，「新町二条下ル町」などのように詳細に記述されているものもある．これらは，人物の活動を特定する上で重要な情報であるといえる．

3.4 人間関係

家族，親，師匠，交友関係，仕えた人・組織の 5 属性を抽出対象とした．親以外の家族に関する記述に対し，親に関する記述が多く，抽出対象として分けている．

3.5 その他人物情報

職業，得意分野，作品，名前が出ている資料の 5 属性を抽出対象とした．得意分野とは職業以外で携わっていたとされる学問および芸術活動を指すが，職業と内容が重複している場合もある．また，官位や法位は職業に含めた．

4. 属性値の自動推定

参考テキスト中に部分文字列として含まれる属性の認識・抽出を自然言語処理における固有表現認識（Named Entity Recognition）のタスクとして定式化する．ラベリング付与形式は BIO 形式 [2] とし，固有表現抽出モデルとしては，深層学習モデルである BiLSTM-CRF [3] を用いる．

4.1 固有表現認識

固有表現認識は与えられたテキストから人名や場所あるいは，時間などを表す文字列もしくは単語列を抽出する自然言語処理のタスクである．提案手法では，文字を単位とする．これは，対象分野が特殊であり，高精度の単語分割器が利用可能ではないと考えられるからである．

固有表現認識は系列ラベリングのタスクとして捉えることができる．系列とは記号の列のことを指し，自然言語において文を構成する単語の列が系列に相当する．系列ラベリングとは，系列に対してラベルづけを行うタスクのことをいう．固有表現認識は，各文字に対して固有表現の一部分か否かのラベルを推定するタスクと定式化できる．

4.2 BIO 形式

BIO 形式は系列ラベリングのタスクで頻繁に用いられるアノテーション形式である．抽出対象となる文字列の始まりを表す B（Beginning），抽出対象となる文字列の継続を表す I（Intermediate），抽出対象ではないことを表す O（Outside）を用いて属性を付与した．

例えばある文中の「漢学者」という文字列に職業の属性（ラベル: Occ）を付与する際，まず事前に文を文字単位に分割しておき，文字列の始まりである「漢」にはラベル Occ-B，それに続く文字列である「学」と「者」にはラベル Occ-I を付与する．

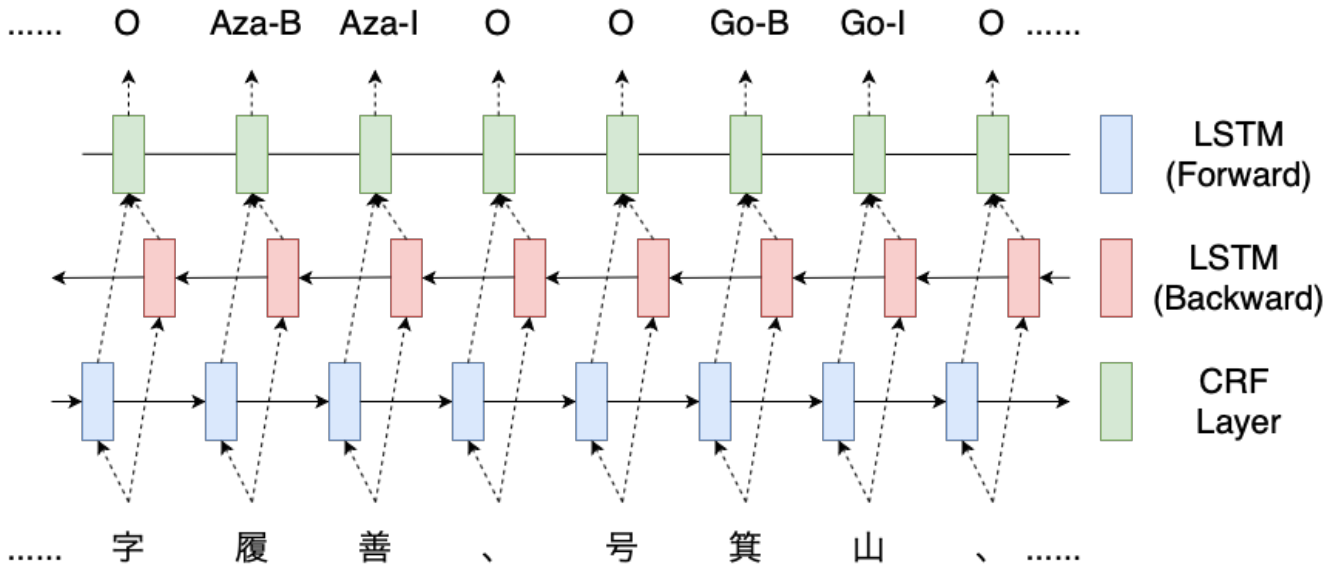

図 1 BiLSTM-CRF モデル．ラベル「Aza」は属性「字」を，ラベル「Go」は属性「号」を表している．

表 2 属性ごとの抽出対象の例．括弧内はアノテーション済みデータにおける出現回数を表す．

属性（出現回数）	高頻度の文字列の例（出現回数）
人名	
姓（57）	源（18），藤原（3），平（3）
名（77）	教文（2），景雷（2），呉春（2）
字（126）	文仲（2），子行（2），節夫（1）
号（223）	沖澹（3），東海（3），雪下庵（2）
その他別名（204）	直徳（2），節助（1），河内主膳（1）
時間情報	
生まれた年（74）	一七三四（2），寛保元年（2），一七三八（2）
没年（285）	天保 2 年（5），天保 10 年（5），一八一八（4）
享年（114）	六十九（8），六十三（6），未詳（5）
空間情報	
出身地（169）	京都（128），江戸（4），浪華（2）
住所（205）	京（20），京都（17），江戸（6）
墓地（57）	百万遍（3），延年寺（2），中堂寺（2）
人間関係	
家族（25）	荷田春満の弟（1），義兄和庵（1），木下順庵の後裔（1）
親（58）	並河友之進宗武の男（1），羽倉信名の養子（1），上林久兵衛の子（1）
師匠（114）	小沢芦庵（8），香川景樹（6），皆川淇園（3）
交友関係（39）	頼山陽（4），蕪村（2），池大雅（2）
仕えた人・組織（58）	徳大寺家（3），鷹司家（2），妙法院宮（2）
その他人物情報	
職業（391）	漢学者（44），医家（25），歌人（24）
得意分野（135）	書（8），篆刻（8），歌（8）
作品（118）	文画誘掖（2），古詩叢（1），国史略（1）
名前が出ている資料（30）	平安人物志（2），海内偉帖人名録（1），大雲院過去帳（1）

4.3 BiLSTM-CRF

BIO ラベルの推定には，双方向長短期記憶ネットワーク（Bidirectional Long Short-Term Memory; BiLSTM）と条件付き確率場（Conditional Random Fields; CRF）を用いることとする．BiLSTM は固有表現認識を含む系列に対するラベル推定のタスクにおいて高い精度を実現することで知られている．CRF は，一貫性のあるラベル列の推定に優れている．これらをモデルのレベルで多段に組み合わせた

BiLSTM-CRF は，固有表現認識の一般的なモデルである．

4.3.1 BiLSTM

BiLSTM とは，入力文字列に対して順方向と逆方向にそれぞれ異なる LSTM を用いることで双方向の情報を単語レベルで抽出するモジュールである．

言語処理における LSTM は，単語をベクトルに変換した単語分散表現を用いている．入力単語列を単語分割して 1 つの単語を 1 つのユニットとし，ユニットごとに処理をしていく．LSTM の特徴はユニットごとの処理の際に，そのユニットより前のユニットまでの情報を記憶するベクトルを用いることである．つまり，処理している単語よりも前にある単語の情報を用いることができる．さらに BiLSTM は LSTM を文頭から文末，および文末から文頭の 2 つの方向に用いることで，ある単語の前だけでなく，後ろにある単語の情報も固有表現の推定に利用することができるようになり，より精度の高い推定が可能となる．

なお，前述のように，高精度の単語分割器が利用可能ではないので，単位を単語から文字に変更している．

4.3.2 CRF

CRF は，文レベルで整合性の高いラベルの系列を推定するためのモジュールである．ある文字列に対して考えられるラベルの列に対しスコア付けを行う．これによって，BiLSTM から整合性がないラベル列が候補として出力された時に，それを排除するように機能する．例えば BIO 形式では，O の直後に I が続くラベル列は解釈不可能である．そのようなラベルの列に対して低いスコアを与えるよう学習することでそのようなラベルの列が出力されないようになる．

5. 評価

提案する属性・属性値の定義およびその自動推定手法を評価するために，人手によるアノテーションとそれを用いた自動推定モデルの構築を行った．本節では，これらについて詳述する．

5.1 人手によるアノテーション

まず，「平安人物志短冊帖」に付されている人物についての説明文 708 件のうち，210 件について人手でアノテーションを行なった．アノテーションした文の数は 1165 文となり，属性を付与した文字列の数は全部で 2559 件となった．属性ごとの文字列の出現回数と各属性における出現頻度の多い文字列の例を**表 2** に示す．文章は単語単位ではなく文字単位で分割した．

5.2 自動推定モデルの構築

次に，人手によるアノテーション結果を学習データとし，自動推定モデルの構築を行なった．

5.2.1 ハイパーパラメータ

BiLSTM の層数は順方向と逆方向ともに 1 層とし，隠れ層の次元数は 320 に設定した．語彙は学習データ中に現れる文字のうち，頻度が 2 以上のものを選択した．学習時のミニバッチサイズは 10 とした．最適化手法には Adam [6] を採用し，初期学習率は 1.0×10^{-3} に設定した．学習時には 500 イテレーション毎に学習データから分割した開発データ上で評価を行い，その精度が前回の評価時から悪化する度に学習率を半減させた．学習率を 3 回半減させた時点で学習を終了し，開発データにおける精度が最もよいパラメータを保存した．

5.3 自動推定実験

属性ごとの抽出のしやすさを比較するため，単一の属性のみを付与した学習データをそれぞれ用意した．評価の際にはデータの偏りによる精度のばらつきの影響を小さくするため，アノテーションされたデータを 7 分割し，学習データ:開発データ:テストデータ = 5:1:1 に分けたデータ群を 7 セット用意してモデルを 7 つ用意し，それぞれのモデルでテストを行った．各属性の結果を**表 3** に，学習データサイズを全体の 1/8， 1/4， 1/2， 1/1 とした場合の結果を**図 2** に示す．

5.4 評価基準

属性値推定モデルの評価には，固有表現認識と同様に，適合率（Precision），再現率（Recall），F 値（F-measure）を用いることとした．それぞれの定義は以下のとおりである．

$$\text{適合率} = \frac{\text{正解であった属性値の数}}{\text{モデルが抽出した属性値の数}}$$

$$\text{再現率} = \frac{\text{正解であった属性値の数}}{\text{テストデータ中の属性値の数}}$$

$$\text{F 値} = \left(\frac{\text{適合率}^{-1} + \text{再現率}^{-1}}{2} \right)^{-1}$$

適合率は，抽出誤りの少なさを表す一方， 再現率は抽出漏れの少なさを表す．これらはトレードオフの関係にあるので，総合的な精度を測るために，これらの調和平均である F 値を用いる．

5.5 全体の結果

全体の結果を表 3 に示す．この表から，属性により精度のばらつきが大きいことがわかる．また図 2 の学習曲線から，ほとんど全ての属性において学習データが増加するにつれて，F 値が向上していること，多くの属性について大幅に向上することがわかる．これは追加の学習データを用意することでさらなる精度の改善が期待できることを示している．

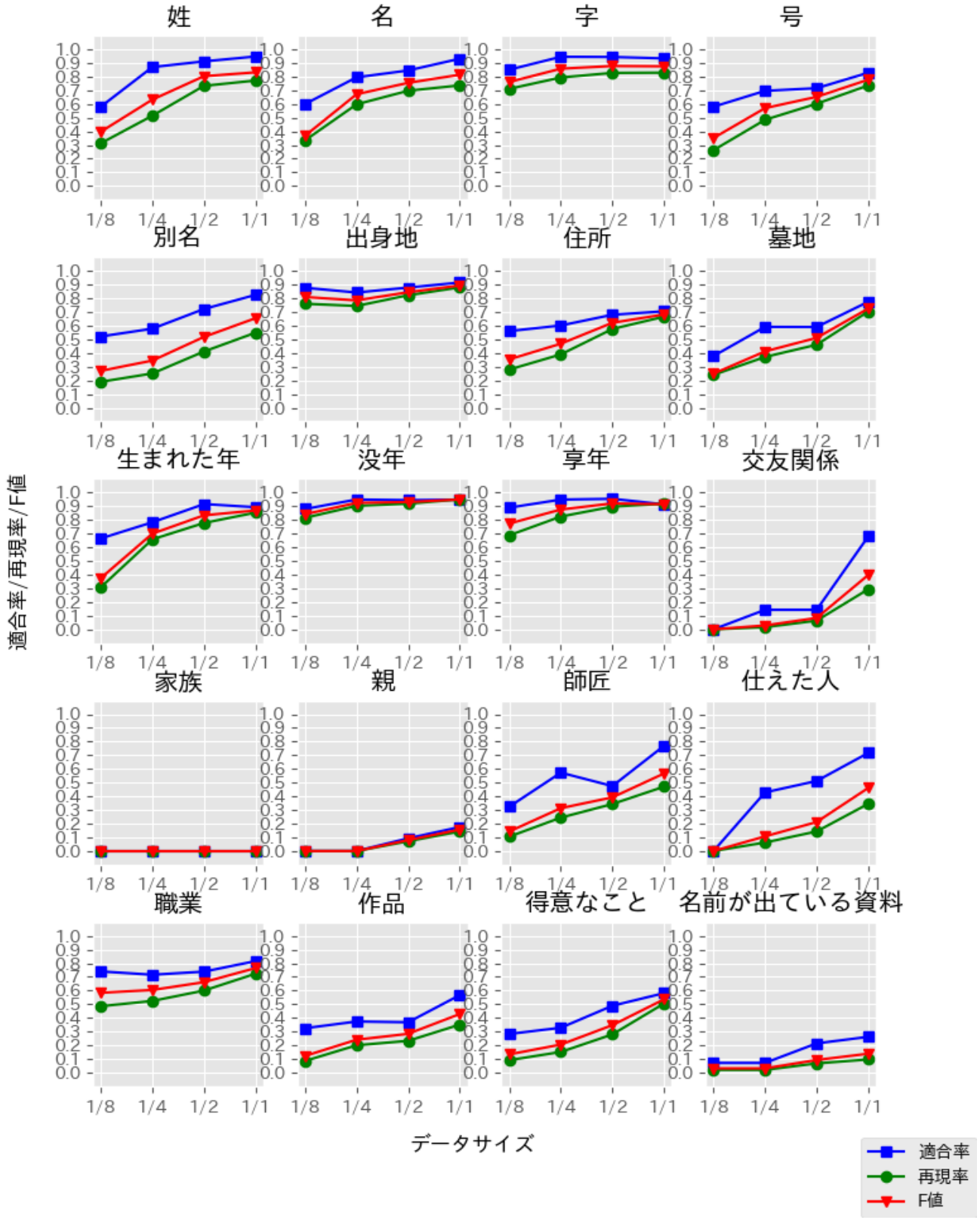

図 2 学習曲線．図中の青線，緑線，赤線はそれぞれ適合率，再現率，F 値に対応している．

表 3 実験結果．各属性における適合率，再現率， F 値を示している．括弧内はアノテーション済みデータにおける出現回数を表す．

属性	適合率	再現率	F 値
人名			
姓（57）	0.9464	0.7707	0.8306
名（77）	0.9299	0.7353	0.8130
字（126）	0.9326	0.8282	0.8748
号（223）	0.8277	0.7347	0.7769
別名（204）	0.8241	0.5479	0.6538
時間情報			
生まれた年（74）	0.8928	0.8530	0.8687
没年（285）	0.9481	0.9501	0.9490
享年（114）	0.9134	0.9177	0.9146
空間情報			
出身地（169）	0.9145	0.8791	0.8926
住所（205）	0.7042	0.6648	0.6809
墓地（57）	0.7738	0.7026	0.7240
人間関係			
家族（25）	0.0000	0.0000	0.0000
親（58）	0.1750	0.1428	0.1568
師匠（114）	0.7685	0.4700	0.5664
交友関係（39）	0.6785	0.2929	0.3971
仕えた人・組織（58）	0.7166	0.3447	0.4607
その他人物情報			
職業（391）	0.8155	0.7232	0.7645
作品（118）	0.5668	0.3511	0.4285
得意なこと（135）	0.5812	0.5053	0.5358
名前が出ている資料（30）	0.2618	0.0965	0.1388

5.6 属性ごとの結果

5.6.1 名前

別名を除き，80%前後の F 値となっている．姓，名，字，号は抽出対象となる文字列が短く，前後の文字列からの推測が容易であると考えられる．別名は長い文字列である場合があること，前後の文字列にバリエーションがあることが抽出を困難にしている要因であると考えられる．

5.6.2 時間情報

どの属性も比較的高い F 値となっている．没年と享年に関しては 90%を超えており，生まれた年に関してもそれに近い値を実現できている．いずれの属性も出現回数が多い上に，前後の文字列からの推測がしやすいことが要因であると考えられる．

5.6.3 空間情報

出身地については 90%に近い F 値となっているが，住所と墓地については 70%前後の値となっている．住所は抽出対象となる文字列が長いものが多いため，抽出が困難になっていると考えられる．

5.6.4 人間関係

全体的に低い F 値となっている．抽出対象となる単語列の出現回数が他の属性に比べて少ないものが多く，データ量の少なさが原因の 1 つとなっていると考えられる．

表 4 誤抽出の例．下線は誤って抽出した文字列．

属性	内容
短冊番号	516
人物名	能勢晴臣
人物情報テキスト	能勢春臣 () **歌人**。姓は**源**。 **竹廼屋**と号し、**竹屋角左衛門**と称した。 **京都**の人。**佛光寺油小路東**に住し、 **和歌**をよくし**古筆刀剣の鑑定**を以て 知られた。
号	竹廼屋，竹，廼屋
住所	佛光寺油小路東， 佛，光寺油小路東
別名	竹屋角左衛門，竹屋角左衛
姓	源
出身地	京都
職業	歌人
得意分野	和歌，古筆刀剣の鑑

表 5 抽出漏れの例．下線は抽出できなかった文字列．

属性	内容
短冊番号	220
人物名	西村仙斎
人物情報テキスト	西村仙斎 () **歌人**。**京都**の人、 住所は人物志に御影堂境内地と出ている。 短冊の書ぶりが余斎翁 (上田秋成) を 学んだもののようである。秋成の門人か。
出身地	京都
職業	歌人，御
師匠	斎翁

5.6.5 その他人物情報

属性によって F 値にばらつきがみられる．職業については F 値が 70%を超えており，他の 3 属性と比べて 20 ポイント以上も大きい値となっている．これは抽出対象の出現回数が多いことや「漢学者」や「医家」など一部の文字列の出現回数が多いことが抽出をしやすくしている要因であると思われる．作品，名前が出ている資料に関しては抽出対象となる文字列が長いこと，得意分野に関しては前後の文脈のパターンが多いことが大きな原因になっていると考えられる．職業に関しては，官位の抽出が文字列の長さゆえに失敗していることがある．

6. 未作業箇所に対する推定

作成したモデルを用いて，アノテーションされていない残りのデータの推定を行った．属性によって抽出の精度の差はあれど，全体的によく抽出できているといえる．

次に，推定に失敗した例をあげる．

6.1 推定に失敗した例

6.1.1 誤抽出の例

表 4 にあるテキストでは，誤って抽出している文字列がある．号や別名，住所については正解の文字列の他に，それらの一部が候補として抽出されていることがわかる．

誤った抽出例は正解の文字列の部分文字列であることが多く，その場合は人手による修正により対処可能である．

6.1.2 抽出漏れの例

表 5 にあるテキストでは，住所である「御影堂境内地」，師匠である「秋成」が抽出できていないことがわかる．住所について，他のテキストでは「御影堂境内地に住し」といったような文脈で出現する割合が圧倒的に多く，表 5 にあるような文脈での出現回数は少ないために抽出漏れが起こっていると考えられる．

7. おわりに

本論文では，平安人物志を題材にして，人物の説明文から各属性の属性値を自動推定することを提案した．そのために，まずできる限り多くの情報を抽出することを目的として 21 の属性を策定した．次に，属性・属性値のアノテーションの基準を定め，一部の説明文について人手でアノテーションを行った．最後に，人手によるアノテーション結果から学習する自動推定モデルを提案し，自動推定実験を行い，評価した．

自動推定結果から，半数弱の属性は現状の少量のアノテーションでも比較的高い推定精度（75%以上）となっているが，残りの属性に関しては大きな精度向上の余地があると考えられる．人手によりアノテーションした量は少ないので，今後の見通しを得るために，学習データのサイズを変化させた場合の精度を測定した．その結果，精度が不十分な属性のほとんどにおいて増量することで精度向上が見込めることが確認された．

今後の取り組みとして，精度向上と並行して抽出された情報を活用することが考えられる．現時点でも比較的高い精度となっている時間に関する記述（生まれた年，没年，享年）と場所に関する記述（出身地，住所，墓地）の活用が有望であろう．これらにより，歴史上の人物がどの時代にどこで活動していたか，つまり時空間に関する情報を計算機で扱うことが可能となる．これにより，歴史学の研究をより円滑に行うことが可能となると考えられる．提案手法は，属性の策定とアノテーションという人手による作業を含むが，「平安人物志」以外のデータにも適用可能であるので，資料横断的に歴史上の人物の時空間に関する情報活用するうえで大きな役割を果たしうる．

参考文献

[1] 白井 圭佑, 森 信介, 後藤 真: 人名辞典からの知識抽出, 人文科学とコンピュータシンポジウム (2020).
[2] Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, Third Workshop on Very Large Corpora (1995).
[3] Guillaume, L., Miguel, B., Sandeep, S., Kazuya, K. and Chris, D.: Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).
[4] Graves, A. and Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural networks, Vol. 18, No. 5-6, pp. 602–610 (2005).
[5] Lafferty, J. D., McCallum, A. and Pereira, F. C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001).
[6] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, International Conference on Learning Representations (2014).