情報処理学会論文誌 Vol.63 No.2 293–301 (Feb. 2022)

人名辞典からの知識抽出

白井 圭佑[1,a)] 松崎 真里[2,b)] 森 信介[3,c)] 後藤 真[4,d)]

受付日 2021年5月17日, 採録日 2021年11月2日

概要：人名辞典等人文学に関わる辞書類からの知識抽出は，それらを用いて人文学研究のための基盤を構築することが可能になるという点で意義がある．一方で，人手による抽出作業は場合によっては高コストになりうる．そこで，本研究では機械学習手法を用いた自動抽出器の構築を試みる．実験結果から，固有表現認識器を用いた芳賀日本人名辞典からの知識抽出では，全体的に高い抽出精度が実現可能であることが分かった．

キーワード：固有表現認識，知識抽出

Knowledge Extraction from a Biographical Dictionary

Keisuke Shirai[1,a)] Masato Matsuzaki[2,b)] Shinsuke Mori[3,c)] Makoto Goto[4,d)]

Received: May 17, 2021, Accepted: November 2, 2021

Abstract: Extracting knowledge from dictionaries like biographical ones is beneficial as they enable us to build a foundation for digital humanities research. However, knowledge extraction from dictionaries by human effort can be costly on large-scale dictionaries. To alleviate this, we developed a method for automatic knowledge extraction and tested it on a biographical dictionary. From experimental results, we found that a named entity recognizer can achieve high accuracy on a Japanese biographical dictionary.

Keywords: named entity recognition, knowledge extraction

1. はじめに

本研究では，人文学の基本的な情報基盤となりうる人名辞典について，機械学習手法を用いた自動抽出器の構築を試みる．人文情報学とデジタルアーカイブの進展により，人文社会系の研究資源データは充実しつつある．そのようななか，これらのデータを基にした様々な横断検索への試みがなされつつあるとともに，これらのデータ群を効果的に結びつけるための，基本的な基盤データの構築が求められるようになってきた．人間文化研究機構においては，歴史地名辞書等の過去の地名情報のデータが公開されている[*1]とともに，関野樹氏を中心として時間情報の統語的な環境が提供される等[*2]，基盤情報は充実しつつあるが，まだ多くの基盤情報が不足している状態であるといわざるをえない．これらの情報の基本となるものの多くは，これまでに作成されてきた紙ベースの辞書類である．これまで長期にわたり，人間が情報を理解するために蓄積されてきた情報を有効活用することで，機械による人文学解析も当然進展する．しかし，人間が理解してきた辞書をもとに，機械的な処理を可能にするためには，一定の変換が必要なの

1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
2 京都大学工学部地球工学科
Undergraduate School of Civil, Environmental and Resources Engineering, Kyoto University, Kyoto 606–8501, Japan
3 京都大学学術情報メディアセンター
Academic Center of Computing and Media Studies, Kyoto University, Kyoto 606–8501, Japan
4 国立歴史民俗博物館
National Museum of Japanese History, Sakura, Chiba 285–8502, Japan
a) shirai.keisuke.64x@st.kyoto-u.ac.jp
b) matsuzaki.masato.86x@st.kyoto-u.ac.jp
c) forest@i.kyoto-u.ac.jp
d) m-goto@rekihaku.ac.jp

*1 歴史地名データ：https://www.nihu.jp/ja/publication/source_map
*2 HuTime: http://www.hutime.jp/

も事実である．特に辞書の中から一定の関係性を抽出し，機械可読の形式に変更することは必須でありながらも，人によるアノテーション付与等が求められることになるため，高コストであり課題となっていた．そこで，本研究では，これを解決するための手段として，機械学習手法による自動抽出器を構築し用いることとした．アノテーション済みの人名辞典を用いて深層学習ベースの固有表現抽出器の構築を試み，これまでの辞書類の蓄積を，新たな情報基盤へ転化するための手法に向けた検討の一歩とする．

2. 関連研究

人物（Person）に関する知識（Knowledge）のテキストからの自動抽出は，その知識を Person-Relation-Knowledge の形で抽出することであり，情報抽出における関係抽出のタスクととらえることができる．ここで，関係（Relation）はその知識が所属するクラスである．特に，対象の人物（Person）の表現が既知であるという設定では，知識とその関係のみが抽出対象となる．この枠組みでは，Wikipedia 等の人物記事における家族関係等の推定を系列ラベリングを用いて行う手法が提案されている [1], [2]．

人物の表現が既知であるという設定は，本研究で対象とする人名辞典の知識抽出においても同様であり，この方向ではパターン学習ベースの手法 [3] や既存の固有表現認識器を用いた手法 [4] が提案されている．しかし，パターン学習ベースの手法では抽出対象のパターンを網羅しきれない可能性があることが，既存のツールを用いる手法では対象とするテキストの言語や抽出したい知識のクラスによってはそのまま適応できないことが，それぞれ問題としてあげられる．

3. 課題

本研究では，与えられた人名辞典に対して，自動抽出器を用いて知識の自動抽出を行う．本章では，まず，3.1 節で人名辞典からの知識抽出の定式化を行う．次に，3.2 節で本研究で用いる人名辞典についての解説を行う．最後に，3.3 節で自動抽出手法について説明する．

3.1 問題の定式化

まず，人名辞典中のある人物に対して，人物を識別する固有 ID d，人物名 p，その解説文 s の三つ組 (d, p, s) が与えられていると仮定する．ここで，解説文 $s = w_1 w_2 \cdots w_n$ は自動抽出機の入力であり，n 個の単語からなる．本研究の目的は，この解説文 s 中に部分文字列として存在する知識を，自動抽出器を用いて属性ごとに抽出することである．ここで，抽出対象の属性の種類数を m としたとき，期待する自動抽出器の出力 y は知識のクラス（以降，属性）a と知識を表す部分文字列（以降，属性値）v の組の集合 $y = \{(a_1, v_1), (a_2, v_2), \cdots, (a_m, v_m)\}$ と表現できる．ここで，属性 a_i（$1 \leq i \leq m$）の属性値 v_i は解説文 s 中の部分文字列であり，属性によっては解説文 s 中に複数個の抽出対象が存在する可能性がある．また，後述するとおり，人物によっては，解説文中に抽出対象の属性値が存在しない例もある．

3.2 人名辞典

本研究では，人名辞典として芳賀矢一著『日本人名辞典』[*3]（以下，芳賀人名辞典）を用いた．芳賀矢一（1867–1927）は，戦前における文学者であり多くの文学テキストの校訂を行うとともに多数の辞書類を作成している．本辞典はそれらの辞書類のうちの 1 つである．この辞典には約 50,000 人の人名とその解説文が収録されている．本辞書は古いものであるものの，以下の利点があると考え，今回の採用に至った．

(1) Public Domain であり著作権処理上の問題がないこと
(2) 網羅的であり Wikipedia 等の既存の辞書にいまだに存在しない人物も書かれていること
(3) 親子関係等の記述が詳細であるとともに，定型的な表現で書かれていること
(4) 著作やどの勅撰和歌集に歌が残されているか等の記述（「作歌」と本辞典上では記載）が詳しく書かれていること
(5) 上記の理由から，抽出したデータをもとに関連したリンクを作ることで，様々な人文系データの基盤となりうるため，人間文化研究機構でデータ基盤として用いる方向で検討を進めていること

我々は，27,059 人分のアノテーション済みデータを保有しており，各人名データには 17 種類の属性を付与している．属性によっては 1 つの属性値をとるものもあれば，複数の値をとるものも存在する．ここで，属性はアノテータによって事前に定義されたものである．属性に対応する文字列は解説文から抽出されたか，あるいはアノテータによって付与されたものである．**表 1** は，人物「鶯笠」における属性と属性値の例である．ここで，解説文中の太字の文字列は各属性における属性値である．また，表には属性「親」と「著作」に対応する値が欠落しているが，これは対応する属性値が解説文中に存在しないためである．

本研究では，抽出対象の属性として，「芳賀人名解説 2」，「人名解説」，「別名」，「親」，「所属組織」，「仕えた人」，「時代」，「死没年月日」，「死没時齢」，「著作」，「作歌」の 11 種類の属性を選択した[*4]．その他の属性は抽出対象外としたが，これはサンプル数が非常に少ないために，事前実験において固有表現認識器の十分な精度が得られなかったため

[*3] 画像 http://www.let.osaka-u.ac.jp/~okajima/kensaku/hagayaiti/
[*4] ここで，「芳賀人名解説 2」はその人物がどういった人物かを端的に述べたものであり，「人名解説」はそれをより記述的に書いたものである．

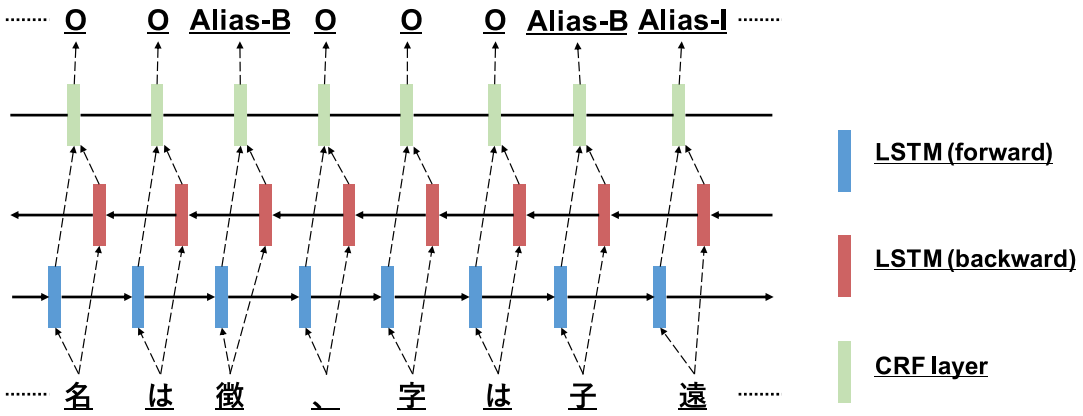


図 1 BiLSTM-CRF モデル．ここで，NE タグ「Alias」は属性「別名」を表している
Fig. 1 Main architecture of the BiLSTM-CRF model. Here, an NE tag "Alias" corresponds to the attribute "Betsumei."

表 1 人物「鷺笠」における解説文と属性の例．解説文中の太字は抽出対象の知識である．また，この例では「親」と「著作」に対応する知識が解説文中に存在しない
Table 1 An example of the person "Ohritsu." Bold strings are the extraction target. In this example, the knowledge for "parent" and "writing" do not exist in the text.

解説文	**東京の俳人。春湖**の門、**塩坪氏、無事庵**と号す。**明治二十七年一月十七日**大阪に歿す。**年七十六**。
属性	属性値
芳賀人名解説 2	東京の俳人
別名	塩坪氏，無事庵
親	——
仕えた人	春湖
死没年月日	明治二十七年一月十七日
死没時齢	年七十六
著作	——

である．また，表 1 でも見られるように，人物によっては上の 11 種類の属性に対応する属性値が存在しないこともある．

3.3 自動認識手法

本研究では，人名辞典の解説文中に部分文字列として含まれる知識の認識・抽出を系列ラベリングとして解く．これには，自然言語処理における固有表現認識器（以下，NER）を用いる．NER の学習には教師信号としてアノテーション済みデータが必要であるが，今回はアノテーション済みの芳賀人名辞典が利用可能であるため，これを用いて NER のパラメータ推定が可能である．NER モデルとしては，深層学習モデルである BiLSTM-CRF [5] を用いる．BiLSTM-CRF は固有表現認識を含む系列ラベリングのタスクにおいて高い精度を実現することで知られており，双方向長短期記憶ネットワーク（Bidirectional Long Short-Term Memory; BiLSTM）[6] と条件付き確率場（Conditional Random Field; CRF）[7] の 2 つのモジュールから構成されるモデルである．ここで，BiLSTM 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り，CRF はラベルの系列を文レベルで推定するためのモジュールである．また，本研究ではラベル列は BIO 形式 [8] で付与する．BIO 形式は系列ラベリングのタスクでよく用いられるタグ付与形式であり，属性値となる文字列の始まりを表す B（Beginning），その継続を表す I（Intermediate），それ以外であることを表す O（Outside）を用いてタグの付与を行う．**図 1** に BiLSTM-CRF モデルの構造を示す．

事前処理として，アノテーション済みデータとして利用可能な 27,059 件のうち，解説文が他の人物への参照となっているものを除いた．その後，前処理後のデータを学習データ，開発データ，テストデータとして分割した．さらに学習データのサンプルから，同じ属性どうしでアノテーション区間が交差するものを取り除いた．これらの処理の結果，各分割におけるサンプル数は，学習データで 17,342 件，開発データで 2,172 件，テストデータで 2,168 となった．**表 2** に各分割における各属性の属性値の出現回数とその種類数を示す．表から，「芳賀人名解説 2」や「時代」，「死没時齢」等の属性では多くの文字列が表層的に一致していることが出現回数と種類数を比較することから分かる．同様に，「人名解説」や「親」，「著作」等の属性では表層的に異なる文字列が比較的多くを占めていることが，出現回数と種類数の比から見て取れる．参考までに，各属性における出現頻度の高い文字列の例を**表 3** に示す．

モデルパラメータとしては，BiLSTM の層数は順方向と逆方向の LSTM ともに 1 層とし，隠れ層の次元数は 576 に設定した．埋め込み層の次元数は隠れ層の次元数と同一の 576 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，その結果 3,009 語になった．学習時のミニバッチサイズは 10 とした．最適化手法には

表 2 各分割における抽出対象文字列の出現回数と種類数
Table 2 The frequency and the number of types of extraction target strings on each split.

属性	学習データ		開発データ		テストデータ	
	出現回数	種類数	出現回数	種類数	出現回数	種類数
芳賀人名解説 2	13,292	5,010	1,658	954	1,661	897
人名解説	6,091	4,736	766	675	752	679
別名	14,443	10,953	1,810	1,646	1,816	1,622
親	5,222	3,945	685	640	663	624
所属組織	3,786	1,230	456	251	465	249
仕えた人	3,837	2,200	477	372	526	426
時代	3,987	590	485	188	503	187
死没年月日	5,446	3,842	704	625	675	605
死没時齢	3,222	237	389	97	381	95
著作	1,407	1,397	173	173	156	156
作歌	3,984	414	522	96	611	118

表 3 属性ごとの抽出対象の例．括弧内は出現頻度を表す
Table 3 Extraction target examples. The number in parentheses represents the frequency.

属性	高頻度の文字列の例
芳賀人名解説 2	俳人 (884)，歌人 (461)，国学者 (305)，備前長船の刀匠 (252)，美濃関の刀匠 (246)
人名解説	京都に歿す (189)，安政年間江戸に住す (83)，京都の人 (68)，江戸の人 (68)，蕉風 (56)
別名	藤原氏 (321)，源氏 (133)，三郎 (82)，太郎 (76)，平氏 (70)
親	後水尾天皇 (20)，景行天皇 (17)，重長 (16)，桓武天皇 (14)，国光 (13)
所属組織	歩兵少尉 (275)，歩兵中尉 (271)，歩兵大尉 (186)，左衛門尉 (108)
仕えた人	本居大平 (167)，本居宣長 (133)，芭蕉 (116)，本居春庭 (75)，蓼太 (49)
時代	元禄 (305)，応永 (196)，天文 (192)，寛文 (166)，文化 (140)
死没年月日	明治三十八年三月 (95)，元治元年 (51)，明治三十七年十月 (50)
死没時齢	年六十九 (108)，年六十五 (93)，年六十六 (93)，年六十一 (91)，年六十二 (90)
著作	詩集 (3)，文集 (3)，医方大成論抄 (2)，文集等 (2)，論語集説 (2)
作歌	新千載 (256)，続千載 (225)，玉葉 (217)，新拾遺 (205)，続後拾遺 (200)

Adam [9] 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 イテレーションごと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たび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良いパラメータを保存し，評価に用いた．モデルパラメータに関しては，解説文によって異なる属性間で属性値の文字列が交差することがあったため，今回は属性ごとに異なるパラメータを用いて NER モデルの学習を行った．また，実験では 1 つの設定に対して異なる乱数シードを用いて 5 つのモデルを学習し，評価時にはそれらの精度の算術平均と標準偏差を報告する．これは BiLSTM-CRF が機械学習モデルであり，疑似乱数の初期値によって精度が変化する可能性があるためである．

4. 評価

4.1 NER による知識抽出の実験と評価

NER モデルの評価には精度（Precision），再現率（Recall），F 値（F-measure）を用いた．**表 4** にその結果を示す．これより，学習後の NER モデルの F 値は 67%から 96%程度であり，属性によって精度にばらつきが存在するものの，全体的には比較的高い抽出精度を実現していることが分かる．特に，「時代」や「死没年月日」，「死没時齢」の F 値は 95%以上であり，これらの抽出精度は実用に適うものだと思われる．「時代」や「死没時齢」，「作歌」に関しては，抽出対象のデータ数や種類数が他属性と比べて少ないにもかかわらず，いずれも 90%以上の F 値を実現していることが分かるが，これはこれらの属性の対象文字列のパターンが比較的推定しやすいためであると考えられる．同様に，「人名解説」や「所属組織」に関しては他の属性と比較して多少精度が劣るが，これは人物によって属性値のパターンが大きく異なり，対象の文字列が推定しにくいためだと考えられる．また，「著作」はその属性値のほとんどが表層的に異なる文字列であるが，実験結果では 75%の F 値をテストデータにおいて実現できていることが分かる．これは対象の文字列が前後の文脈から推定しやすいためだと

表 4 実験結果．各属性における精度，再現率，F 値を平均と標準偏差とともに示している
Table 4 Experimental results. Precision, recall, and F-measure were reported with mean and standard deviation.

属性	精度	再現率	F 値
芳賀人名解説 2	0.9077 (± 0.0076)	0.9040 (± 0.0135)	0.9057 (± 0.0047)
人名解説	0.6996 (± 0.0252)	0.6543 (± 0.0274)	0.6752 (± 0.0077)
別名	0.8970 (± 0.0064)	0.9196 (± 0.0048)	0.9081 (± 0.0023)
親	0.9371 (± 0.0093)	0.9430 (± 0.0018)	0.9400 (± 0.0048)
所属組織	0.7404 (± 0.0230)	0.6916 (± 0.0183)	0.7147 (± 0.0105)
仕えた人	0.8733 (± 0.0173)	0.8202 (± 0.0092)	0.8458 (± 0.0088)
時代	0.9462 (± 0.0095)	0.9563 (± 0.0013)	0.9512 (± 0.0050)
死没年月日	0.9506 (± 0.0037)	0.9570 (± 0.0032)	0.9538 (± 0.0019)
死没時齢	0.9375 (± 0.0030)	0.9916 (± 0.0020)	0.9638 (± 0.0010)
著作	0.7791 (± 0.0276)	0.7359 (± 0.0282)	0.7567 (± 0.0255)
作歌	0.9515 (± 0.0032)	0.8606 (± 0.0149)	0.9037 (± 0.0083)

表 5 外部辞書を用いた実験結果
Table 5 Experimental results of using an external dictionary.

属性	精度	再現率	F 値
別名	0.8927 (± 0.0069)	0.9181 (± 0.0051)	0.9052 (± 0.0041)
親	0.9358 (± 0.0154)	0.9367 (± 0.0053)	0.9361 (± 0.0074)
仕えた人	0.8830 (± 0.0177)	0.8122 (± 0.0091)	0.8460 (± 0.0103)

考えられる．

次に，**図 2** に抽出対象の 11 属性について，学習データのサイズを 1/8，1/4，1/2，1/1 と変化させた場合の学習曲線を示す．ここで図中の実線は乱数シードを変えた場合のスコアの算術平均を，陰影はその標準偏差を表している．図から，すべての属性において利用可能な学習データ量が増加するにつれて，F 値が向上していることが分かる．特に，「人名解説」，「所属組織」，「仕えた人」，「著作」はこの中でも大幅な精度向上を見せており，追加のアノテーション済み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が分かる．また，「時代」，「死没年月日」，「死没時齢」の属性は学習データが少量の場合でも十分に高い精度を実現していることが分かる．

さらに，追加実験として外部辞書 [10] を用いた実験を行った．実験では，人名表現を対象とし，属性「別名」，「親」，「仕えた人」から抽出した属性値を用いることで擬似的に外部辞書を構築した．また，特徴量は，ある単語を先頭とする連続する n 文字が外部辞書中に存在すれば 1 を，存在しなければ 0 を与えるように bool 値で計算した．今回は，この特徴量を $n = 1, 2, 3, 4$ に対して計算し，4 次元のベクトルとして BiLSTM の入力に結合する形で与えた．**表 5** に実験結果を示す．表 4 の結果と比較すると，外部辞書を用いた場合と用いない場合とでは，NER の抽出精度にはほとんど変化がないことが分かる．これは深層学習以前の NER モデルとは異なる傾向 [10] であるといえるが，近年の研究でも似たような結果が報告されている [11]．外部辞書として，学習データから構築したものではない，より包括的なものを用いて特徴量を計算することで，最終的な NER の精度に変化を与える可能性もあるが，これに関しては今後の課題としたい．

4.2 抽出された知識の利用例

前節まででは，NER を用いた人名辞典からの知識抽出について記述した．抽出した知識の応用に関しては，たとえば，

- 知識の可視化 [12]
- 検索エンジンへの応用[*5]
- 質疑応答システムへの応用 [13]

のようなものが考えられる．本節では，上記の応用例のうち，可視化を行うシステムを実装し解説する．

ここでは，人物間の関係をグラフとして可視化するシステムを実装した．グラフの可視化ツールには `force-graph`[*6] を用いた．グラフの節点には芳賀人名辞典に記載された人物を，辺にはその関係を表す属性を，それぞれ対応させた．今回は属性として，「親」と「仕えた人」を対象とした．また，属性は方向性を持つため，グラフ中のラベルに「<親」のような方向を付与することで関係性の方向を表現した．これに加え，人手アノテーションは実線，NER による自動アノテーションは点線で示した．

例として，**図 3** に仁徳天皇，応神天皇，仲哀天皇，日本

[*5] https://blog.google/products/search/introducing-knowledge-graph-things-not/
[*6] https://github.com/vasturiano/force-graph（2021 年 4 月アクセス）

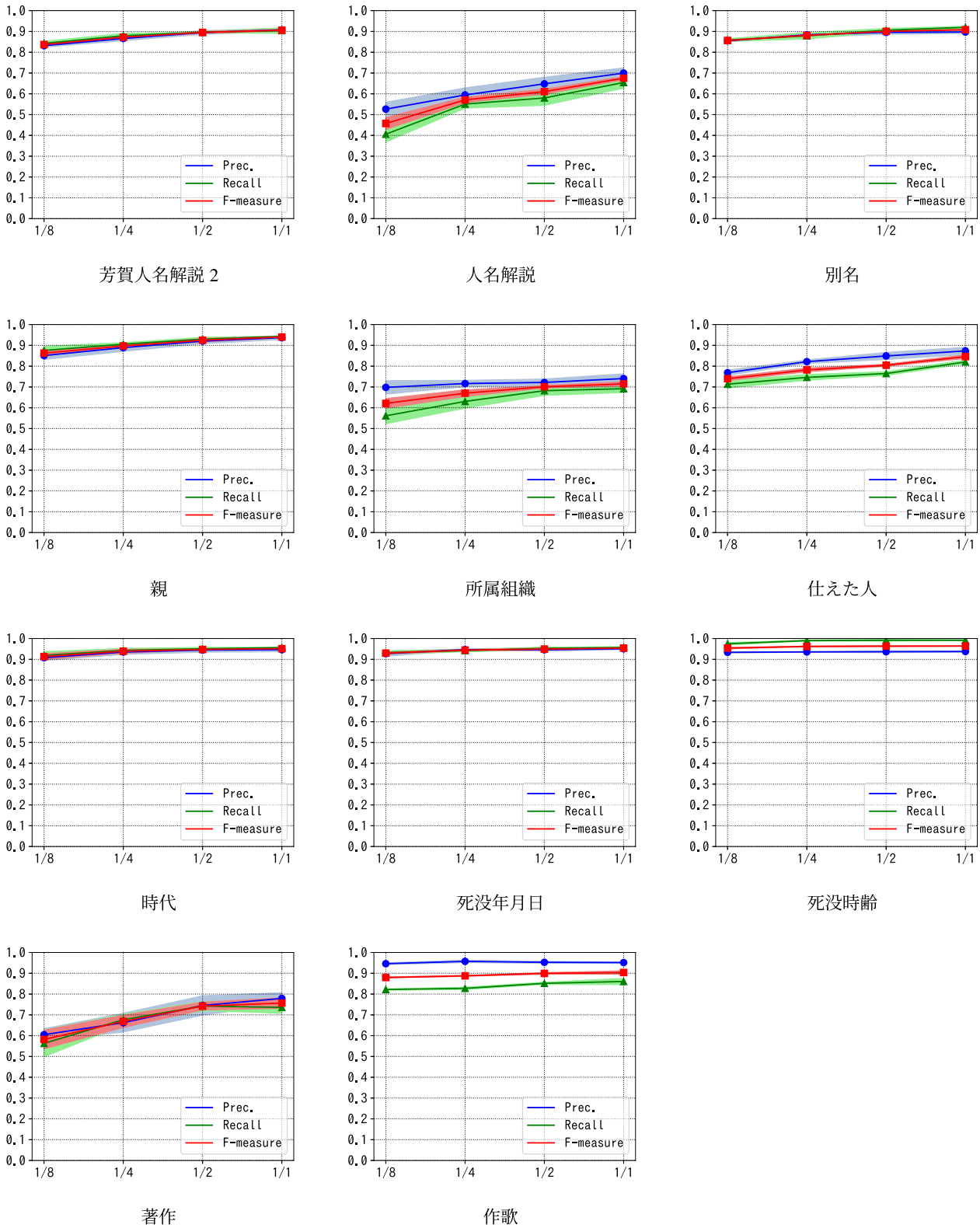


図 2 学習曲線．図中の青線，緑線，赤線はそれぞれ精度（Prec.），再現率（Recall），F 値（F-measure）に対応している．また，実線は 5 つの異なる乱数シードを用いて NER モデルの学習を行った場合のスコアの算術平均を，陰影はその標準偏差を表している

Fig. 2 Learning curves. Blue, green, red lines represent precision, recall, and F-measure, respectively. The solid line represents the arithmetic mean of the NER models trained with five different seeds, and the shade illustrates the standard deviation.

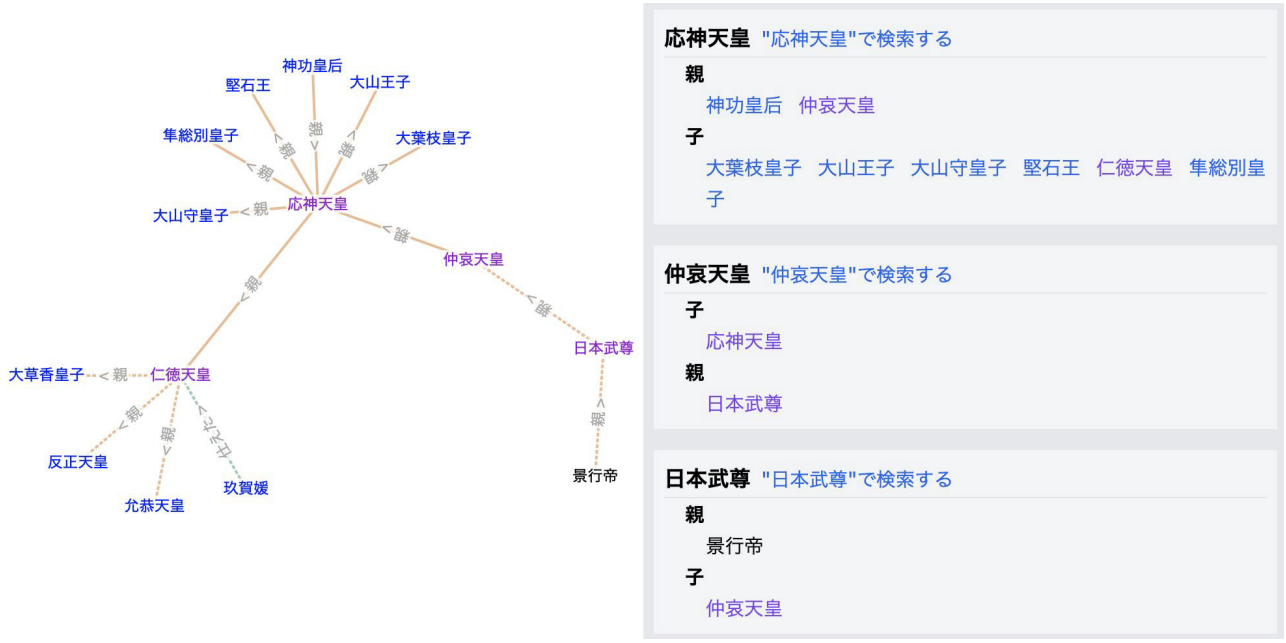


図 3　人物関係のグラフ

Fig. 3　A graph of the relationship between persons.

図 4　人物関係を利用した検索例

Fig. 4　A search example of utilizing the relationship.

武尊周辺の人物関係のグラフを示す．グラフに示された人物関係のうち，仁徳天皇，仲哀天皇，日本武尊に関する人物関係は点線で示されており，これは NER の自動アノテーションにより推定された部分である．自動アノテーションによって芳賀人名辞典に記載された人物関係が補完され，より広い人物関係のつながりを表示できるようになったことが分かる．

人物関係のグラフによる可視化のさらなる応用として，資料検索と連動させて検索機能を拡張することが考えられる．図 3 の右項にある “仁徳天皇で検索する” から，国立歴史民俗博物館学術情報リポジトリ[*7]の資料を検索した例を図 4 に示す．この例では，仁徳天皇に関する資料とともに，その子女に関する資料を関連項目として検索結果に表

[*7] https://rekihaku.repo.nii.ac.jp（2021 年 4 月アクセス）

示している．検索結果のハイライト部分に表示された人名から人物関係のグラフに戻り，その人物を中心とした人物関係を調べることも可能である．これらの機能によって，検索者は，検索対象の人名が記載された資料に加えて，関連のある人物の人名が記載された資料まで存在することを受動的に知ることができる．

ここまで，人物関係の可視化に着目して「親」と「仕えた人」という 2 つの属性に絞って紹介したが，その他の属性，たとえば「著作」や「作歌」等といった属性に対しても同様に資料検索機能の拡張に応用できると考えられる．

5. おわりに

本研究では，辞書類からの知識抽出を自動的に行うために，深層学習ベースの固有表現認識器を用いた．アノテーション済みの芳賀人名辞典を用いた実験では，対象とした属性に対して全体的に高い抽出精度を実現していたほか，一部の属性に対してはデータの追加によりさらなる精度改善が見込めることも実験的に確認した．今後としては，2 つの方向性が考えられる．まず，今回の実験に用いた芳賀人名辞典に関して，アノテーションされていない残りの人名データに対する属性の自動抽出および人手による確認作業である．実験結果から示したとおり，NER モデルは過半数の属性において高い抽出精度を実現しており，これは残りの人名データにおける人手抽出作業の効率化に大きく貢献するものと考えられる．次に，人名辞典から抽出した知識を基にした知識グラフの構築である．これに関しては，「時代」や「死没年月日」等の一部属性における表現の正規化を行ったうえで，既存のオントロジや知識ベースとの各知識のエンティティ・リンキングを行うことで段階的に構築を目指したい．

参考文献

[1] Mann, G. and Yarowsky, D.: Multi-Field Information Extraction and Cross-Document Fusion, *Proc. 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL '05*), pp.483–490, Association for Computational Linguistics (2005).

[2] Culotta, A., McCallum, A. and Betz, J.: Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text, *Proc. Human Language Technology Conference of the NAACL, Main Conference*, pp.296–303, Association for Computational Linguistics (2006).

[3] Nikesh, G. and David, Y.: Structural, Transitive and Latent Models for Biographic Fact Extraction, *EACL*, pp.300–308, Association for Computational Linguistics (2009).

[4] Samet, A. and Vincent, L.: A comparison of named entity recognition tools applied to biographical texts, *ICSCS*, pp.228–233, IEEE (2013).

[5] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[6] 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).

[7] Lafferty, J.D., McCallum, A. and Pereira, F.C.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. 18th International Conference on Machine Learning*, pp.282–289 (2001).

[8] Ramshaw, L. and Marcus, M.: Text Chunking using Transformation-Based Learning, *3rd Workshop on Very Large Corpora* (1995).

[9] Kingma, D.P. and Ba, J.: Adam: A method for stochastic optimization, *International Conference on Learning Representations* (2014).

[10] Ratinov, L. and Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition, *Proc. 13th Conference on Computational Natural Language Learning* (*CoNLL-2009*), pp.147–155, Association for Computational Linguistics (2009).

[11] Magnolini, S., Piccioni, V., Balaraman, V., Guerini, M. and Magnini, B.: How to Use Gazetteers for Entity Recognition with Neural Models, *Proc. 5th Workshop on Semantic Deep Learning* (*SemDeep-5*), pp.40–49, Association for Computational Linguistics (2019).

[12] Tamper, M., Hyvönen, E. and Leskinen, P.: Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research, *CICling*, Springer-Verlag (2019).

[13] Hao, Y., Zhang, Y., Liu, K., He, S., Liu, Z., Wu, H. and Zhao, J.: An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, *Proc. 55th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), pp.221–231, Association for Computational Linguistics (2017).

白井 圭佑

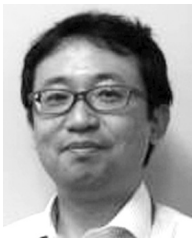
京都大学大学院情報学研究科博士課程在学中．

松崎 真里

京都大学工学部地球工学科在学中．

森 信介 （正会員）

1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム（株）入社．2007 年京都大学学術情報メディアセンター准教授．2016 年より同教授．京都大学博士（工学），音声言語処理および自然言語処理に関する研究に従事．1997 年情報処理学会山下記念研究賞受賞．2010 年，2013 年情報処理学会論文賞受賞．2010 年第 58 回電気科学技術奨励賞．言語処理学会，ACL 各会員．



後藤 真 （正会員）

2007 年大阪市立大学大学院文学研究科後期博士課程修了．同年日本学術振興会特別研究員（PD）．2008 年花園大学文学部専任講師．2015 年国立歴史博物館准教授．博士（文学）．人文情報学・歴史情報学を専門とする．2003 年情報処理学会山下記念研究賞受賞．