人名辞典からの知識抽出

白井 圭佑 (京都大学 情報学研究科)
森 信介 (京都大学 学術情報メディアセンター, Linfer)
後藤 真 (国立歴史民俗博物館)

人名辞典など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為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える. 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.

Knowledge Extraction from a Biographical Dictionary

Keisuke Shirai (Graduate School of Informatics, Kyoto University)
Shinsuke Mori (Academic Center for Computing and Media Studies, Kyoto University & Linfer)
Makoto Goto (Research Department, National Museum of Japanese History)

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. In this paper we describe the method and report experimental results.

1 はじめに

本発表では, 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ものである. 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. そのような中, 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, 歴史地名辞書など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など [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言わざるを得ない. 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など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.

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

2 関連研究

辞書類からの知識の自動抽出を行うための手段として, テンプレートを用いたパターンベースの手法 [1] や既存の固有表現認識ツールの利用 [2] が考えられる. しかし, 前者は対象によってはパターンを網羅することが困難となる可能性があること, 後者は抽出対象とするテキストの言語によっては対応していない場合があることが問題として挙げられる. 本研究では, 深層学習ベースの固有表現認識器を用いる.

参考テキスト	**画僧**。名は**真瑞**、**紀州の人**。画を**野呂介石**に学び、遂に一家を成す。

属性	属性値
芳賀人名解説 2	画僧
人名解説	紀州の人
別名	真瑞
仕えた人	野呂介石

表 1: 人物「愛石」における参考テキストと他属性の抽出対象の文字列の例.

3 課題

人名辞典には人物に関する解説文（以下, 参考テキスト）が付与されており, これには関連人物の名称やその人物に関わる時代等の様々な属性の情報が部分文字列として含まれている. これらの人手による抽出作業は, 対象とするデータの規模が大きい場合には高コストになる. また, 対象となる属性の抽出には一定の専門性が必要となる．クラウドソーシング等による効率化も考えられうるが，本研究では, これらの人文知識情報の機械による効率的な抽出方法の開発という観点からも事前にアノテーション済みの辞書類を用いて固有表現認識器を学習し, 抽出に用いることでこれらの問題に対処することとした.

3.1 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一つである．この辞典には約 50,000 人の人名とその参考テキストが収録されている. 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. 親子関係など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などの記述（「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. 属性によっては一つの属性値を取るものもあれば, 複数の値を取るものも存在する. ここで, 属性はアノテーター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参考テキストから抽出されたか, あるいはアノテーターによって付与されたものである. 例として, 表 1に人物「愛石」の属性を示す. ここで, 参考テキスト中の太字の文字列は各属性における属性値である.

本研究では属性値が参考テキスト中に部分文字列として存在するという仮定を置いた. 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. 本研究ではその他の属性を抽出対象外としたが, これはサンプル数が非常に少ない為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である. また, 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. 例えば表 1では「親」や「所属組織」等の属性値が参考テキスト中に含まれていない.

4 自動認識手法

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のタスクとして捉える. 固有表現認識器 (以下, NER) の学習には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データとして芳賀人名辞典が利用可能であるため, これを用いた 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 [3] を用いる. BiLSTM-CRF は固有表現認識を含む系列ラベリング

[3]画像 http://www.let.osaka-u.ac.jp/~okajima/kensaku/hagayaiti/.

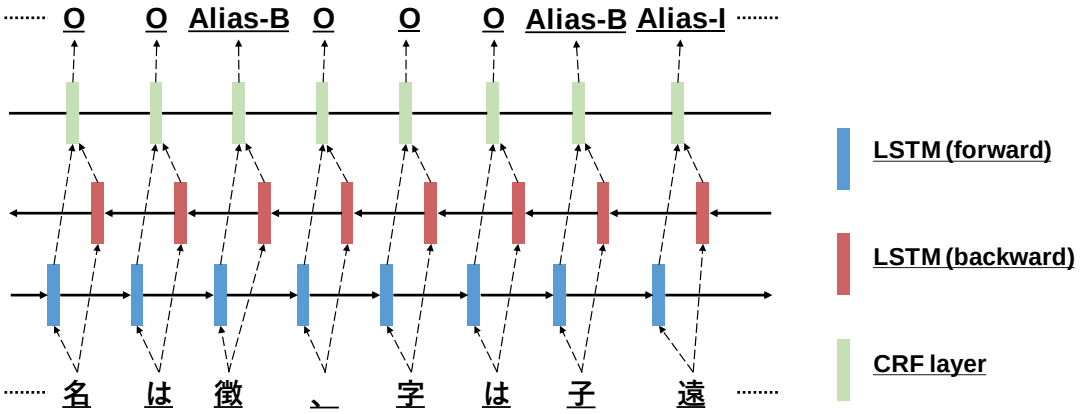


図 1: BiLSTM-CRF モデル. ここで, NER タグ「Alias」は属性「別名」を表している.

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)

表 2: 属性毎の抽出対象の例. 括弧内は出現頻度を表す.

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) [4] と条件付き確率場 (Conditional Random Field; CRF) [5] の 2 つのモジュールから構成されるモデルである. 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る. 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 [6] 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. 1に BiLSTM-CRF モデルの構造を示す.

学習時には, アノテーション済みデータとして利用可能な 27, 059 件の内, 参考テキスト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同士でアノテーション区間が交差するものを取り除いた. これらの処理により, 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 である. 表 3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わかる. 同様に, 「人

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

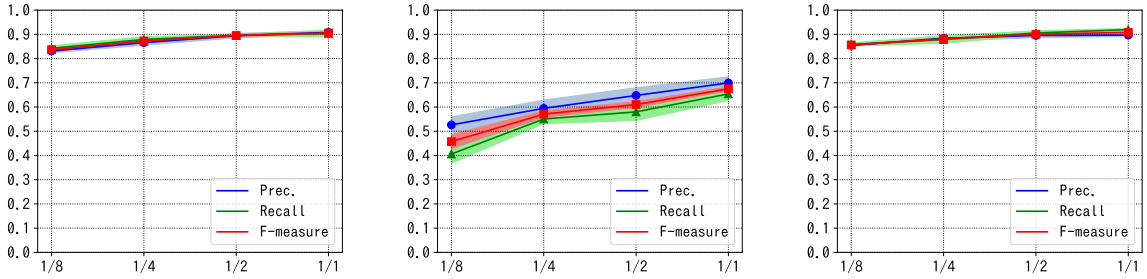
表 3: 各分割における抽出対象文字列の出現回数と種類数.

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)

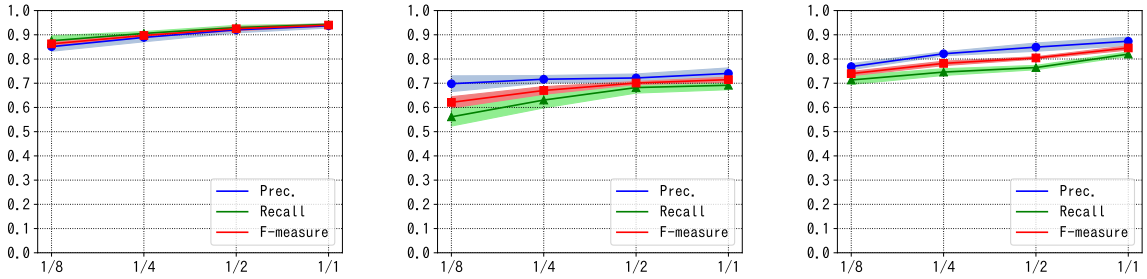
表 4: 実験結果. 各属性における精度, 再現率, F 値を平均と標準偏差と共に示している.

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 2に示す.

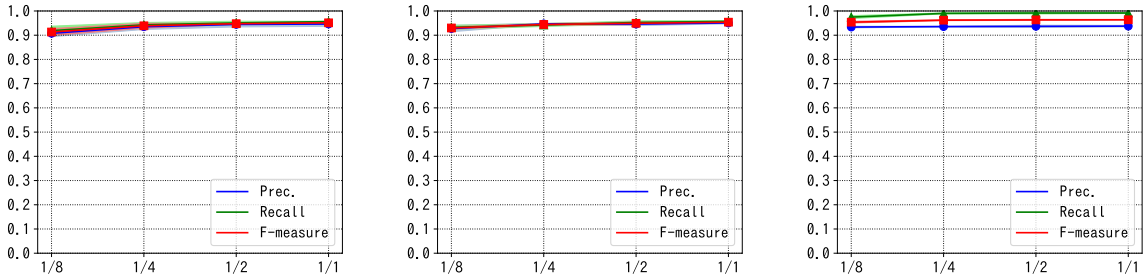
モデルパラメータとしては, BiLSTM の層数は順方向と逆方向の LSTM 共に 1 層とし, 隠れ層の次元数は 576 に設定した. 埋め込み層の次元数は隠れ層の次元数と同一の 576 に設定した. 語彙は学習データ中に現れる文字のうち, 頻度が 2 以上のものを選択し, その結果 3,009 語になった. 学習時のミニバッチサイズは 10 とした. 最適化手法には Adam [7] を採用し, 初期学習率は 1.0×10^{-3} に設定した. 学習時には 500 イテレーション毎に学習データから分割した開発データ上で評価を行い, その精度が前回の評価時から悪化したする度に学習率を半減させた. 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. モデルパラメータに関しては, 参考テキスト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, 実験では一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為である.



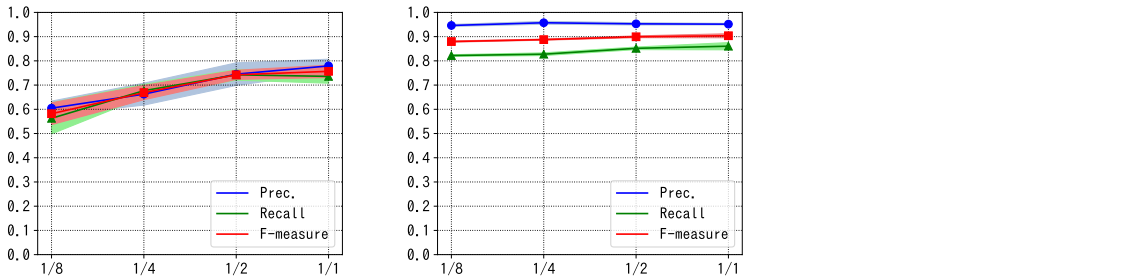
芳賀人名解説 2　人名解説　別名



親　所属組織　仕えた人



時代　死没年月日　死没時齢



著作　作歌

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.

5 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わかる. 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関わらず, いずれも 90% 以上の F 値を実現していることがわかるが, 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為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為だと考えられる. また, 「著作」はその属性値の殆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出来ていることがわかる. これは対象の文字列が前後の文脈から推定しやすい為だと考えられる.

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, 全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わかる. 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出来ることがわかる. 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わかる.

6 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及び人手による確認作業である. 実験結果から示した通り, 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上で, 既存のオントロジー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] Nikesh, G. and David, Y.: Structural, Transitive and Latent Models for Biographic Fact Extraction, *EACL*, Athens, Greece, Association for Computational Linguistics, pp. 300–308 (2009).

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

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

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