

事例ベース固有表現抽出を用いた古文からの歴史人物情報の抽出および活用

苑 広媛¹ 李 康穎² 後藤 真³ 木村 文則⁴ 前田 亮⁵

¹立命館大学 情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1

²立命館大学 総合科学技術研究機構 〒525-8577 滋賀県草津市野路東 1-1-1

³国立歴史民俗博物館 〒285-8502 千葉県佐倉市城内町 1-1-7

⁴尾道市立大学 経済情報学部 〒722-8506 広島県尾道市久山田町 1600 番地 2

⁵立命館大学 情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1

E-mail: gr0580hp@ed.ritsumei.ac.jp kangying@gst.ritsumei.ac.jp m-goto@rekihaku.ac.jp

f-kimura@onomichi-u.ac.jp amaeda@is.ritsumei.ac.jp

あらまし 近年、人文学の研究に情報技術を応用する人文情報学の研究が進展している。たとえば辞書類から歴史上の人物の情報を抽出することで、歴史人物の研究や分析を行うことが可能になるなど、自然言語処理技術を用いることで、大量のデータを効率的かつ迅速に処理することができると考えられる。一方で、機械学習を用いた情報抽出においては、学習データの作成に手作業による多大な時間を要する問題がある。そこで、本研究では、few-shot 学習によって日本人辞典から歴史的人物情報を抽出する手法を提案し、異なる BERT 事前学習モデルの性能比較を行い、その結果を報告する。さらに、日本人辞典の解説文を用いた人物関係の可視化への活用を試みる。

キーワード 日本人辞典, 固有表現抽出, Few-shot 学習

1. まえがき

近年、情報技術の急速な発展に伴い、インターネットやコンピュータを使った情報収集やデータ処理が普及し、人文学の研究においても情報技術を応用する人文情報学の研究がますます進展している。このような環境下で、人文情報学のさらなる発展を目指すためには、基盤情報の整備が重要になっている。図書館などの情報機関は長年、多くの歴史的文献を保存し、人文学研究のための重要な基盤を提供している。最近の自然言語処理技術の発展に伴い、情報抽出技術を用いることで、重要な情報および豊富なデータを人文学研究に提供することができる。特に、人文学関連辞書から情報を抽出することで、歴史上の人物の研究や分析に利用することができ、人文学研究の基盤を確立することができる。しかし、現在でもまだ多くの基盤整備が不足している状況である。このような状況では、少ないデータを対象とするアルゴリズムが非常に重要である。few-shot 学習 (few-shot learning : 以下 FSL) とは、少ないデータ (画像やテキストなど) で効率的に学習することができ、過去に学習した経験値を応用し、新しいクラスの学習を追加することができるようになる学習手法である。

そこで、本研究では、few-shot 学習に基づく事例ベース固有表現抽出モデルにより『日本人辞典』から歴史的人物情報を抽出し、モデルに用いられる異なる BERT 系事前学習モデルの性能を比較することを目的

とする。具体的には、日本語版 Wikipedia および青空文庫のコーパスを用いて構築された複数の DeBERTa, RoBERTa, BERT などの事前学習済みモデルの性能を比較した結果を分析することで、モデル精度向上の方向性を検討し、さらに抽出結果を用いて人物関係図などの可視化への応用を試みる。

本研究では、人名辞典として芳賀矢一著『日本人辞典』を研究対象とする。この辞書は、古代から明治までの約五万余名の人名を収録し、簡潔に記述したものである。著者である芳賀矢一は戦前の文学者として活躍し、数多くの古典文学の校訂や辞書の作成などを行った人物である。図 1 に、本研究で使用する芳賀矢一著『日本人辞典』の原文の一部を示す。本研究で使用するデータは、『日本人辞典』の原文をテキスト化したものの一部である。抽出する人物情報の例として、人物「足利義詮」の解説文と、そこから抽出すべき属性を図 2 に示す。



図 1 『日本人辞典』(出典：国立国会図書館、請求記号 281.03-H117n)

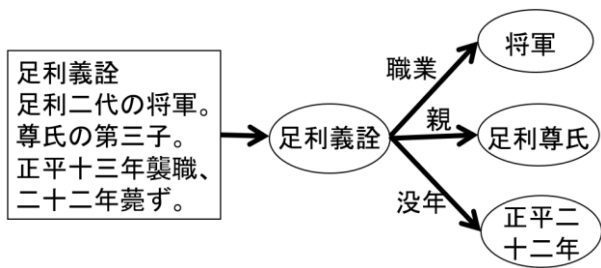


図 2 人物「足利義詮」の解説文と、そこから抽出すべき属性の例

2. 関連研究

固有表現抽出 (Named Entity Recognition: 以下 NER) は、テキストから重要な情報を抽出するための、自然言語処理において重要なタスクの一つである。このタスクの目的は、文章中に出現する固有名詞 (人名, 地名, 組織名など) などの固有表現を自動的に抽出することである。本研究では、デジタルテキスト化された『日本人名辞典』の解説文から人物の属性情報を抽出するための情報抽出手法として、NER を利用する。NER の手法には、パターンベースの手法や伝統的な機械学習に基づく手法が存在する[1]。また、自然言語処理分野で深層学習が普及するにつれて、これに基づく手法もますます普及してきている。例えば、白井ら[2]は、BiLSTM-CRF モデルを用いて本研究と同じ『日本人名辞典』から人物属性を抽出する研究を行い、結果として全体的に高い精度が得られた。

異なる事前学習モデルを比較する研究として、神田ら[3]は、著者推定タスクにおいて複数のデータソースで訓練された BERT モデルの学習済み重みの性能比較を行い、異なる学習データがそれぞれのタスクに与える影響を分析した。鈴木ら[4]は、決算短信や有価証券報告書を用い、金融ドメインのタスクにおいて汎用コーパスを用いた BERT と ELECTRA モデルの性能を比較した。本研究では、FSL に基づく固有表現抽出器を使用し、さまざまな事前学習モデルの性能を比較評価する。

3. 提案手法

本研究では、事例ベース固有表現抽出 (Example-based NER) モデル[5]に基づき、BERT 系の事前学習モデルを使用する。本研究で使用するモデルを図 3 に示す。

本研究のモデルは、2 つの部分に分けられる。一つ目の部分では、与えられた文内にある固有表現の可能性が最も高い範囲の開始位置と終了位置を識別する。クエリに対応するサポートが作成され、固有表現の範

囲を示すタグとして、<e>や</e>などの特殊文字が生成される。サポート内の固有表現の前後にタグが追加される。二つの BERT 系エンコーダ構造を使用してクエリとサポートサンプルをそれぞれエンコードし、サポートサンプル内の <e>と</e>とクエリ内の各文字単位のトークンの一致する位置を計算し、クエリ内の固有表現の開始位置と終了位置を計算する。二つ目の部分では、抽出された固有表現の範囲情報に基づいて、この固有表現がどの種類の固有表現に対応するかをさらに識別する。サポートサンプルを使用し、クエリ内の各トークンが各固有表現の種類の開始位置と終了位置である確率を計算する。つまり、本モデルは、固有表現の範囲を識別する部分と、これに基づいてデータを分類する部分の二つに分けられる。

このモデルにある二つの BERT 構造は、クエリ Encoder とサポート Encoder として用いられる。学習用のラベル付きデータの学習および固有表現の抽出を行いたい新たな解説文の処理に対する学習のそれぞれに対して、サポート Encoder とクエリ Encoder の別々の二つの BERT 構造を用いて学習を行う。サポート Encoder は、ラベル付きデータから分散表現を抽出したものであり、固有表現の出現位置の計算とカテゴリ推測のために用いられる。クエリ Encoder は、ラベルなしの解説文から分散表現を抽出することができる。

抽出されたクエリベクトルとサポートベクトルは、文レベルのアテンションを組み込んだトークンレベルの類似度の計算により固有表現の分散表現を学習する。学習済みモデルは、新しく与えられた少ないラベル付きテキストのサポート例から、クエリ文書の固有表現の情報を予測することができる。

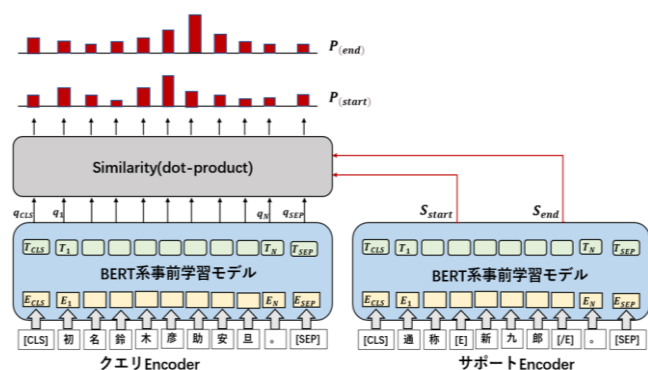


図 3 本研究で用いる Example-based NER モデル

事前学習モデルについては、安岡[6]による、青空文庫 DeBERTa モデルによる国語研長単位係り受け解析の研究で使用されたモデルから選択した。単語単位の事前学習モデルとして、DeBERTa モデルがある。具体的には①deberta-base-japanese-aozora, ②deberta-large-

japanese-aozora, ③deberta-base-japanese-wikipedia および④deberta-large-japanese-wikipedia が含まれる。文字単位の事前学習モデルとして、RoBERTa および BERT の二つのモデルがある。具体的には⑤roberta-base-japanese-aozora-char, ⑥ roberta-large-japanese-aozora-char, ⑦bert-base-japanese-char-v2 が含まれる。

4. 実験

4.1 データ分析

本実験では、専門家により人物に関する属性が付与された解説文のうち、1,717 件を対象として分析を行った。付与された属性として、「地理情報」、「別名」、「所属組織」、「時代」、「著作」、「作歌」、「親」、「妻/夫」、「子供」、「仕えた人」、「死没年月日」、「死没時齢」の 12 種類の属性がある。

これらのうち「地理情報」などの属性は、全人物の解説文において異なる人物の解説に複数回現れる可能性があるという特徴を持つため、本研究では以下の二つのラベリング方法を使ってデータにラベル付与して実験を行った。具体的には、1) 専門家が各人物にラベルを付けた属性だけを用いて人物ごとの解説文にラベリングを行う方法、2) 解説文全体に現れた「地理情報」などの属性を整理して全人物の解説文をマッチングして一致する固有名詞にラベリングを行う方法の 2 種類である。各方法で抽出したラベル数を表 1 に示す。

表 1 共通属性の実験設定およびラベリング数

共通属性	人物ごとの解説文	全人物の解説文
地理情報	167	273
別名	1730	1798
所属組織	300	473
時代	70	76
著作	637	647
作歌	3	13

伝統的な NER 手法では、まず「人物」であることを判別し、全ての人物名を「Person」などの属性として抽出する。さらに抽出された「親子」、「妻/夫」などのエンティティ間の関係を抽出する手法である関係抽出 (relation extraction: RE) を行う。本研究で用いるデータでは、「親」、「妻/夫」、「仕えた人」などの関係が属性として付与されているため、本研究では、人物関係を含む属性と、特定の時間や時代に関連する属性である「死没年月日」、「死没時齢」を、人物ごとの解説文から別々に抽出して実験比較を行う。これらの属性における抽出したラベル数を表 2 に示す。表中のデータは重複を削除した後のものである。

表 2 関係や特定を含む属性のラベリング数

関連性や特定の情報を含む属性	人物ごとの解説文
親	314
妻/夫	39
子供	22
仕えた人	113
死没年月日	844
死没時齢	651

4.2 実験概要と実験結果

本研究では、専門家によってラベル付けされた 1,717 件のテキストデータを用いて、前節で説明した三つのデータセットを構築した。それらは、抽出した共通属性の人物ごとの解説文 (2,907 件)、共通属性の全人物の解説文 (3,280 件)、関連性や特定の情報を含む属性の人物ごとの解説文 (1,983 件) である。各データセットの 20% をテストデータとして使用し、残りの 80% を学習データとして使用して、提案手法の実験を行った。実験結果を表 3 に示す。各モデルの番号に対応するモデル名は、第 3 章で説明したものである。

表 3 提案手法による実験結果

モデル	(共通) 人物ごとの解説文	(共通) 全人物の解説文	(関連性や特定の情報を含む属性) 人物ごとの解説文
①	0.208	0.181	0.595
②	0.159	0.120	0.464
③	0.261	0.217	0.699
④	0.155	0.138	0.357
⑤	0.287	0.227	0.538
⑥	0.211	0.161	0.440
⑦	0.269	0.221	0.636

①と⑤、②と⑥を比較対象にすると、文字単位の事前学習モデルは、単語単位の事前学習モデルよりも優れた結果を得られた。これは、国語研長単位 UD (Universal Dependencies) に基づいて構築された事前学習モデル DeBERTa が形態素解析の手法と類似しているからである。この手法では、まず文を品詞に分割し、解析する処理を行う。古文の形態素への分割においては、抽出精度が低いのが現在の研究での難点である。そのため、文字単位の事前学習モデルの方が、より良い精度が得られている。

5. 実験結果の活用と考察

本研究における実験で得られた歴史人物情報の抽出結果の活用例の一つとして、本章では抽出結果の可視化方法の比較と分析を行う。

5.1 可視化ツールの比較と分析

複雑なデータをわかりやすいグラフに変換すると、データをより理解するのに役立つ。可視化のツールとしては、NetworkX[7], Multiplex[8], pyvis[9], Plotly[10], Visdcc[11], D3Blocks[12] などがある。

NetworkX は複雑なネットワーク構造を作成するために使用できる。ノードとエッジを作成し、ネットワーク分析と可視化を実行し、さまざまなアルゴリズムとグラフィックレイアウトを可視化するための API を提供する。

Multiplex は NetworkX と Matplotlib に基づく複数のネットワーク内のノードとエッジ間の相互作用、および異なる時間と空間での変化を分析するために使用される。そして、生成された画像はより鮮明になる。

pyvis はインタラクティブな可視化とデータレポートを作成するための NetworkX ベースのツールで、ノード、エッジ、さらには完全なレイアウトをカスタマイズするためのさまざまなスタイルオプションも提供する。設定ペインからいつでもさまざまなオプションを Python の辞書として最終設定にエクスポートできる。この辞書は、関数を呼び出すときに config として渡すことができ、ネットワークがそのまま描画される。同時に、折れ線グラフ、ヒストグラム、散布図などのさまざまなグラフの種類とスタイル、およびマウスホバリング、ズーム、パンなどのインタラクティブなコントロールとイベントを提供できる。

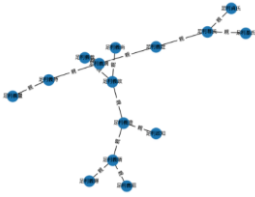


Plotly も NetworkX ベースで Multiplex と比較して Dash フレームワーク上で実行できる。Dash は JavaScript を必要としない分析アプリケーションを構築するためのオープンソースフレームワークで、Plotly グラフィックライブラリと緊密に統合される。同じことが Dash フレームワークと Visdcc で使用でき、通常は Python で構成できる D3Blocks で使用できる。また、Jupyter Notebook での可視化とデータ処理もサポートしている。他のツールと比較すると、よりダイナミックで美しいアニメーション効果が得られる可能性がある。

本研究では、提案手法により抽出した足利家族の人物情報に基づき、上記のツールのうち三つを使用して比較した。実験結果を表 4 に示す。次に、5.2 節でラベル付きの『日本人名辞典』の解説文に基づき、効果の高い pyvis を用いて人物関係の社会ネットワークを分析する。

5.2 可視化手法の分析

人物関係の可視化は時系列によって表すことができる。生年月日や歴史的な出来事などを時系列で可視化することで、人物の生涯や歴史上の位置を明確にすることができる。人名辞典に記載された人物は、通常

表 4 異なる可視化ツールの比較

ツール	例
NetworkX[7]	
Multiplex[8]	
pyvis[9]	

親族関係や社会関係など、特定の社会的な繋がりを持つ。

専門家によるラベル付きの『日本人名辞典』の解説文に基づき、人物間のマッピングを行い、pyvis を使用して人物関係の社会ネットワークを作成する。人物を表すノードに繋がっているエッジが多いほど、より多くの人物がその人物と関連している。また、ノードの影響力は、そのノードの隣接ノードの数と隣接ノードの影響力によって決まる。つまり、高い影響力のある人物は、重要な人物と親密な関係を持っていたり、広い人脈を持っていたりすることが示される。図 4 に『日本人名辞典』の解説文による可視化の例を示す。皇族や有名な家族を中心として展開する社会ネットワークが得られた。皇族のメンバ間の社会的なノードが多く、社会的な関係が複雑であることが示されている。また、皇族の世襲が社会や時代に対してより深い影響を与えている可能性も示唆していると考えられる。

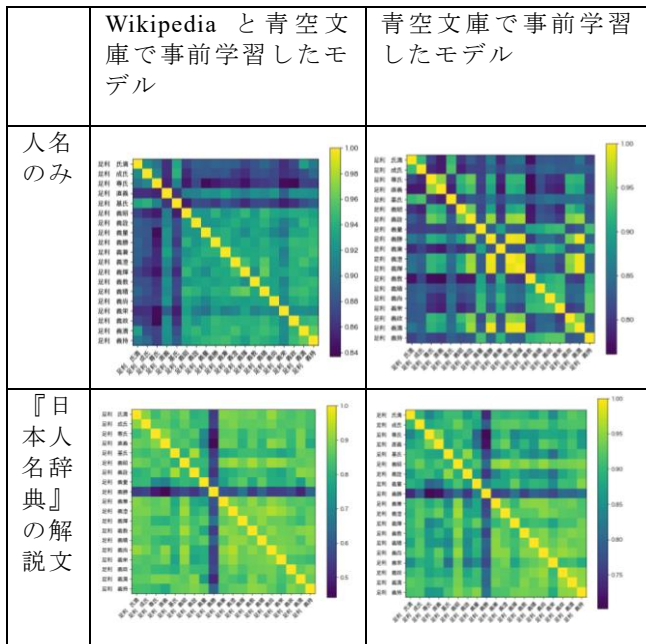
次に、ベクトル化したデータに対して、ヒートマップで人物背景間の関連性の分析を行うことを試みた。ヒートマップは、データの密度を色で表現する可視化ツールである。一般的に、暗い色は低い値を表し、明るい色は高い値を表す。『日本人名辞典』の解説文と人名のデータに対して、それぞれ日本語版 Wikipedia と青空文庫のテキストで事前学習された DeBERTa(V2)モデル [6] と青空文庫のテキストで事前学習された DeBERTa(V2)モデル [13] を使用してベクトル分散表現を抽出する。得られたヒートマップを表 5 に示す。ヒ

ートマップの横軸と縦軸は足利家族の一覧で、明るい部分が多いほど、その部分のデータ値が高い。つまり関連性が高いことを示す。『日本人名辞典』の解説文を使用した場合は、人名のみのデータより高い関連性がある結果が得られ、Wikipedia と青空文庫のテキストで事前学習されたモデルを使用した場合は、より関連性が高い結果が得られた。

Wikipedia のデータを用いた学習済み重みには、豊富な人物情報と人物関係情報が含まれていることは明らかであり、今後の本研究の推進にとって重要な参考となる。また、ベクトルの可視化を使用することで、事前学習モデルの学習データをより直感的に理解することができ、それによりモデルの精度向上とモデル改善に向けた可視化な情報を取得できる。

『日本人名辞典』は日本歴史上の多くの重要人物に関する情報が収録され、不可欠な歴史資料である。日本の文化や歴史などの分野の研究者にとって重要であることは当然ながら、日本の歴史や文化を後世に残すためにも重要な意義がある。可視化技術を使用することで、データをよりよく理解することができ、それにより、学生の学習や研究者の研究を支援できる可能性があると考えられる。

表 5 ベクトル分散表現によるヒートマップ



5.3 最新のツールによる可視化

ChatGPT[14]は OpenAI によって訓練された大規模な言語モデルであり、transformer に基づき、巨大な知識を保有し、GPT-3 モデルを活用した対話システムである[15]。また、言語モデル (Language Model : LM) [16] と強化学習 (Reinforcement Learning : RL) [17]の技術

を使用することで、ChatGPT は人間の言語をより良く理解し、対話応答の能力を向上させている。GraphGPT[18]は、グラフィック情報に基づく言語生成モデルである。グラフィックとその関係に基づき、グラフィック情報を記述する言語を生成することができる。ChatGPT と GraphGPT を組み合わせることで可視化に応用できる可能性がある。ChatGPT は言語情報を処理し、GraphGPT はグラフィック情報を処理する。これにより、より正確で豊富な言語情報を生成することができる。表 6 は、『日本人名辞典』のある人物の解説文を ChatGPT を用いて翻訳した結果である。また、この結果を基に GraphGPT を用いて可視化した結果を図 5 に示す。

大規模データから重要な情報を抽出する研究は重要であると考えられる。既存データを使用して質問を生成し、大規模な生成モデルから既存のデータを拡張する手法が期待される。

表 6 ChatGPT による翻訳の結果

入力データ	数学家。最上流の祖。出羽最上の人。初名鈴木彦助安旦。字は子貴、自在亭と号す。通称算左衛門。江戸に出でて徒に教ふ。天元術に精し。文化十四年十月二十六日没す。年七十一。改精算法等数十部の著あり。
ChatGPT からの英訳出力結果	Mathematician. The best in his field. A person from the highest level in Dewa. Originally named Suzuki Hiko-suke Adan. He had the pen name "Shikan" and called himself "Zuitei." Commonly known as "Susaemon." He lived in Edo and taught mathematics as a hobby. He was proficient in the study of Tengan. He passed away on October 26th, year 14 of the Bunka era at the age of 71. He wrote several works including "Improved Calculation Methods."

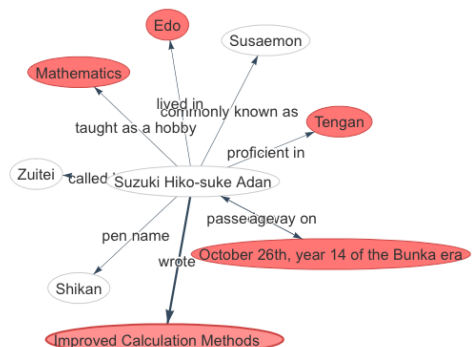


図 5 GraphGPT による人物情報の可視化の例

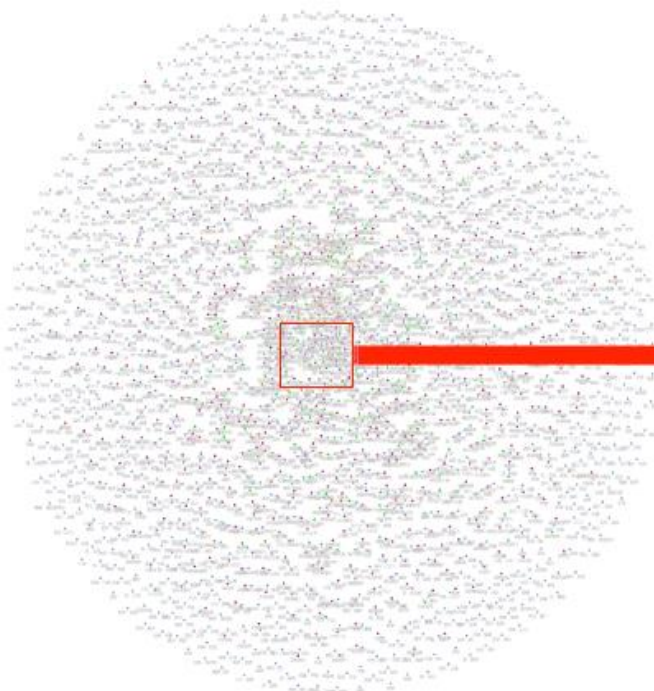


図 4 人物関係ネットワークの可視化例

6. あとがき

本研究では、FSL に基づいて『日本人名辞典』から歴史人物情報を抽出する手法を提案し、異なる事前学習モデルの性能評価のための対照実験を行った。その結果、単語単位の事前学習モデルと比較して、文字単位の事前学習モデルの方がより良い結果を得られた。日本語版 Wikipedia の事前学習モデルを使用すると、モデルの精度が向上することもわかった。また、解説文による人物関係の可視化分析を行った。今後は、抽出結果を活用したクロスドメインのデータからの人物関係抽出などの研究を検討する。

謝辞

本研究の一部は JSPS 科研費 20K12567 の助成を受けたものである。

参考文献

- [1] Samet Atdağ and Vincent Labatut: A comparison of named entity recognition tools applied to biographical texts, Proc. ICSCS 2013, pp. 228–233. 2013.
- [2] 白井圭佑, 松崎真里, 森信介, 後藤真 “人名辞典からの知識抽出” じんもんこん 2020 論文集, 2020.
- [3] 神田泰誠, 柳燐佳, 金明哲 “著者推定における異なる事前学習データを持つ日本語版 BERT の性能比較分析”, 日本行動計量学会第 50 回大会抄録集, pp. 315-316, 2022.

- [4] 鈴木雅弘, 坂地泰紀, 平野正徳, 和泉潔 “事前学習と追加事前学習による金融言語モデルの構築と検証”, 人工知能学会第 28 回金融情報学研究会 (FIN-028), 2022.
- [5] Morteza Ziyadi, Yuting Sun, Abhishek Goswami, Jade Huang, and Weizhu Chen: Example-Based Named Entity Recognition, arXiv:2008.10570 [cs.CL], 2020.
- [6] 安岡孝一 “青空文庫 DeBERTa モデルによる国語研長単位係り受け解析”, 東洋学へのコンピュータ利用 第 35 回研究セミナー予稿集, 2022.
- [7] <https://networkx.org>
- [8] <https://github.com/NicholasMamo/multiplex-plot>
- [9] <https://pyvis.readthedocs.io/en/latest/index.html>
- [10] <https://plotly.com/python/network-graphs/>
- [11] <https://github.com/jimmybow/visdcc>
- [12] <https://github.com/d3blocks/d3blocks>
- [13] 安岡孝一 “Transformers と国語研長単位による日本語係り受け解析モデルの製作”, 人文科学とコンピュータ研究会報告 (CH-128), 2022
- [14] <https://openai.com/blog/chatgpt/>
- [15] Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners, NIPS 2020, pp.1877–1901. 2020.
- [16] Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155, 2022.
- [17] Christiano, P. F., Leike, J., Brown, T., et al.: Deep reinforcement learning from human preferences, NIPS2017, 2017.
- [18] <https://graphgpt.vercel.app/>