情報処理学会論文誌 Vol.63 No.2 310–323 (Feb. 2022)

郷土に残存する江戸期古記録の機械可読化を目的とした市民参加および機械学習による固有表現抽出

吉賀 夏子[1,a)] 堀 良彰[2,b)] 只木 進一[3,c)] 永崎 研宣[4,d)] 伊藤 昭弘[1,e)]

受付日 2021年5月18日, 採録日 2021年11月2日

概要：わが国には，江戸時代以前に記された業務記録や証文などの古記録が数多く存在する．これらを有効に活用するためには，少ない工数で機械可読データを構築する必要がある．特に，地域特有の資料の場合には，地域特有の固有表現への対応が必要となる．本研究では，江戸期の業務日誌である「小城藩日記データベース」の目録記事文から Linked Data などの機械可読データを生成することを具体的目標とし，固有表現抽出の効率化を行う．その第 1 の手法は，市民参加による人手そのものの有効活用である．第 2 の手法は，機械学習による固有表現の自動抽出である．これらの手法を組み合わせることで，通常は収集の難しい地域特有の固有表現を記事文から，自動かつ高精度で抽出可能である．

キーワード：江戸期古記録，シチズンサイエンス，ディープラーニング，固有表現抽出，単語分散表現

Named Entities Extraction by Citizen Participation and Machine Learning for Making Machine-readable Old Records of the Edo Period Remaining in Local Communities

Natsuko Yoshiga[1,a)] Yoshiaki Hori[2,b)] Shin-ichi Tadaki[3,c)] Kiyonori Nagasaki[4,d)] Akihiro Ito[1,e)]

Received: May 18, 2021, Accepted: November 2, 2021

Abstract: There are many ancient documents such as business records and testimonials written before the Edo period in Japan. Machine-readable metadata will be one of effective tools for utilizing those records. In cases of materials related to a very small area, in particular, it is necessary to deal with unique expressions restricted in the area. In this study, we set a specific goal to generate machine-readable metadata such as Linked Data from the database of the cataloged articles for the Ogi-han Nikki (business records) from the Edo period. We aim to improve the efficiency in extraction processes of named entities. For this purpose, we employ two methods. The first is effective use of human resources through citizen participation. The second is automated extraction of named entities by machine learning. We show that the proposed method works well even for materials related to a local area.

Keywords: old local records in Edo period, citizen science, deep learning, named entity extraction, word embeddings

[1] 佐賀大学地域学歴史文化研究センター
The Center for Regional History and Culture, Saga University, Saga 840–8502, Japan
[2] 佐賀大学総合情報基盤センター
Computer and Network Center, Saga University, Saga 840–8502, Japan
[3] 佐賀大学理工学部
Department of Science and Engineering, Saga University, Saga 840–8502, Japan
[4] 一般財団法人人文情報学研究所
International Institute for Digital Humanities, Bunkyo, Tokyo 113–0033, Japan
a) natsukoy@cc.saga-u.ac.jp
b) horiyo@cc.saga-u.ac.jp
c) tadaki@cc.saga-u.ac.jp
d) nagasaki@dhii.jp
e) itouaki@cc.saga-u.ac.jp

1. はじめに

1.1 背景

我々の身近には，江戸時代以前の経済，政治，文化，災害あるいは事件などの出来事における因果関係を明らかにする手がかりとなる古記録が数多く存在している．近年は，古記録を含む様々な文化財をデジタル画像にしておき，書誌情報付きのデータベースとしてWeb公開する取り組みが大量の埋蔵文化財データの利活用，保護対策にも有効な方法として一般化している．それらの画像から記載内容を抽出し，大量の資料を機械的に分析できる形，すなわち機械可読データへ再構成する取り組みも始まっている．たとえば，画像からテキストを抽出する翻刻，TEI（Text Encoding Initiative）[1] や Linked Data [2] 化などによる文書の構造化などである．このような技術の進歩とともに，特に地域に多く散在する古記録が研究組織の枠を超えて利用可能な状況になりつつある．

しかし，大量に残存する地域の古記録を機械可読データに変換するのは容易ではない．なぜなら，各資料の書誌情報作成については，時代や地域，文化，研究分野コンテキストで異なる派生的な情報を十分に含んだ項目とそれらの関係を示す構造化データのモデリング技術が求められる．加えて，可能な限り少ない労力で，古記録画像から記載内容をテキストデータとして抜き出し，多くの人に機械的分析可能な形で提供する工夫も必要である．

筆者はこれまでに，記載内容の機械可読化に係る工数を抑える手法を提案し，実践してきた．具体的には，江戸期小城藩（現在の佐賀県）における業務日誌の目録である「日記目録[*1]」を収録した「小城藩日記データベース [4], [5]」の全73,984件の機械可読化を例に，筆者がこれまでに構築してきたLinked Data自動変換システム [6] を用いて，記事文から**表 1** に示すとおり，人名，地名，出来事名，定型的な候文用語などの重要キーワードとなる固有表現抽出 [7] を行い，最終的にLinked Dataを構築した（**図 1**）．

このLinked Data自動変換システムでは，Web上の例文などから収集した固有表現をユーザ辞書へと保存することで，高精度で形態素解析を実施することができることを示した[*2]．

1.2 課題およびその解決手法

地域特有の固有表現抽出については，Web上の例文から情報を収集すること自体が困難である．特に，人名は資料上に頻回出現するわけではないため，筆者らが先に提案した形態素解析ツールで特定の人名をユーザ辞書に登録して抽出するような手法は，大量の資料に対して有効ではない．

その一方で，日本人の人名を一定数見慣れた人が手作業で人名の抽出を行うと，仮に「七田助右衛門」という人名の存在が明らかであれば，「七田次右衛門」のような類似した文字列に対しても人名であると見当をつけることができる．

本研究では，従来の形態素解析ツール用の辞書に地域の固有表現を蓄積する手法に加え，固有表現抽出において人間ならば容易にできることを機械学習に取り入れるため，1) 研究組織内にとどまらず，シチズンサイエンス（2.1節参照）の枠組みを通じて，郷土資料の読み解きを実際に行える地元市民の協力により，地域の固有表現を抽出した [3], [8]．次に，2) 正解データとして必要な候文の数を可能な限り抑え，多義語に対して文脈を考慮した固有表現クラスの選択を可能にするため，多義語に対応可能な単語分

表 1 日記目録における固有表現クラスの一覧および説明 [9]
Table 1 List and description of named entity classes in the Nikki Mokuroku, summarized lists from Nikki records.

固有表現クラス名	説明
EVENT / 出来事	出来事の名称
TERMS / 候文用語	接続詞，定型句
ROLE / 役職・役割	役職，家族関係
PERSON (JINMEI) / 人名	人名，呼称
PLACE / 場所	地名，建物の呼称
QUANTITY / 数量	数および単位を表す語
DATE / 日時	日時を表す語

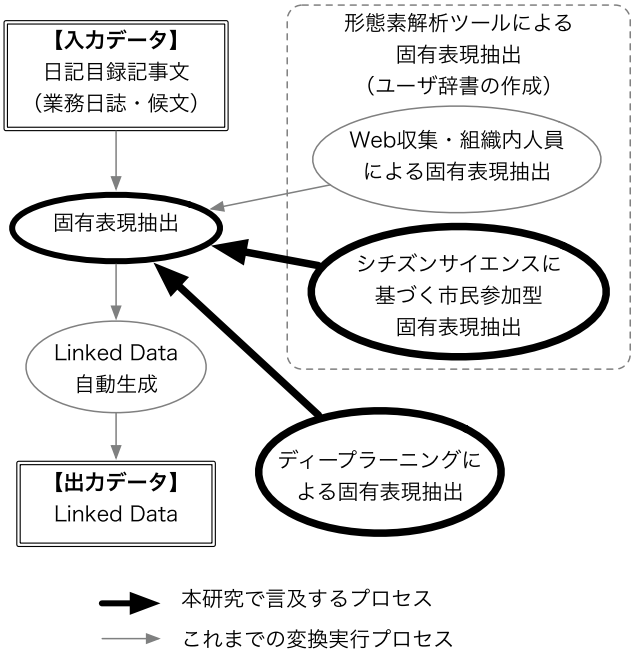


図 1 Linked Data 自動変換システム [6] において本研究が言及するプロセス [8]（太線枠内）
Fig. 1 Points of interest in the automatic Linked Data conversion system (ellipses with thick lines).

[*1] 小城藩の業務日誌は「日記」と呼ばれる．一方，小城藩士が検索の利便のために「日記」の内容を要約して，別途時系列で編集した冊子は「日記目録」という [3].
[*2] 形態素解析ツール MeCab のユーザ辞書を使った具体的手法は https://crch.dl.saga-u.ac.jp/nikki/dataset/に示す．

散表現構築による固有表現抽出手法を採用した．

2) について，具体的には，「単語分散表現」を用いた固有表現抽出分類モデル（以下，判定モデル）を用いて，翻刻テキスト中の固有表現を推測する手法をとる．単語分散表現とは，言語を構成する単語や文の基礎的な意味情報を数値化したベクトル表現データである．近年のディープラーニングによる自然言語処理手法である BERT [10]，Flair [11] などでは，大量の Web 記事コーパスなどからあらかじめ構築した，多義語に対応する単語分散表現を事前学習データとし，別途ドメイン特有のテキストで構築した単語分散表現を追学習用データとして固有表現判定モデルを生成する．

本研究において，我々は，Wikipedia 記事から構築した現代日本語の事前学習データと市民により内容確認済みの候文テキストから作成した追学習データを用いて，候文に対して固有表現を抽出可能な固有表現判定モデルを生成した [9]．

これまでの研究から，固有表現を列挙するだけでも現代日本人に大まかな意味が通じ，形態素解析ツールにおいても現代語の入った基本辞書と候文から収集したユーザ辞書の組合せで高精度な抽出結果を得ている．そのため，我々は，抽出した単語どうしの関係を数値化した単語分散表現では，現代日本語文と江戸期に書かれた候文それぞれの分散表現に大きな違いがないという推測を立てた．

すなわち，前述の現代日本語で作られた単語分散表現データに，分野に特化した候文のそれを追加して学習を行うことで，固有表現の判定モデル構築が可能と判断した．実際には，現代日本語で作られた単語分散表現データ（事前学習データ）は，事前に作成されたものが Web でダウンロード可能である．我々が用意するデータは，固有表現の内容を確認済みの追学習用候文データである．

1) および 2) の取り組みにより，地元の古記録に対し関心の高い市民と研究者による共同作業で，関連地名，人名，行事などの地域特有の固有表現の抽出と追学習用データの作成を行うことができた．

最終的に，前述の事前学習データと追学習用データを基に生成した固有表現判定モデルによる固有表現抽出精度を検証した結果，全記事文 73,984 件の 3 分の 1 である 25,000 件の候文データを用いた判定モデルおよび精度検証において，全固有表現数の 95%を占める人名，候文用語，場所，出来事，役職・役割の各クラスで高精度な結果を得られた．

本稿の構成は，次のとおりである．市民参加による地域の知識と教師データ集積の手法については 2 章，ディープラーニングによる江戸期候文からの固有表現抽出手法については 3 章，4 章では 2 章および 3 章で示した手法を実行した際の固有表現抽出結果および提案手法の有効性について述べる．5 章をまとめとする．

2. 市民参加による郷土知識の収集と固有表現抽出手法

2.1 背景

本研究では，シチズンサイエンスの実践を目的の 1 つとする．学術研究に市民が積極的に参加するシチズンサイエンス [12] は，自然科学などの先行分野だけでなく，近年では人文社会科学分野など多岐にわたる学問分野において，科学者と市民が協働し科学と社会のために新しい知識を生み出す実践手段として注目されている [13]．

わが国の人文情報学および図書館情報学の分野においても，たとえば，手書き文書の翻刻 [14], [15]，画像デジタルアーカイブ構築の素材収集 [16] などで一定の成果をあげている．

本研究では，シチズンサイエンスの実践として，大量かつ地域特有の固有表現抽出タスクを地元市民の支援の下で実施する．同時に，元々地元の歴史や文化に強い関心を持つ参加市民が本タスクを行うことで，参加者がデータベースから様々な語彙とその実践的用法を学び合い，古記録読み解きのスキルや知的探究心の向上を期待できる．

本研究における実践を実現するには，以下にあげる固有表現抽出タスク（以降，タスクと呼ぶ．）を実行する参加者，その支援ソフトウェア，参加者による抽出結果の質について留意し，固有表現抽出の品質を損ねないための工夫を行う必要がある [8], [17]．

2.2 タスク参加者

タスク対象記事文は現代文ではなく「候文」で書かれた日記目録の翻刻済みテキストである．そのため，候文の文法に加えて地域の地名，人名などにもある程度の知識を持つ人物が必要である．抽出内容の質を担保したい場合，タスクを実行可能な人物は限定される．

そこで，候文からの固有表現抽出が可能な専門的人材を探した結果，研究者のみでなく，小城市立歴史資料館の協力により，郷土の知識になじみがあり，候文原文の読解について習熟している 60 代から 80 代の地元市民を中心に固有表現抽出を依頼した．最終的に，記事文 4 万件の固有表現抽出を 8 名で行った．そのうち，3 名から無償，5 名から有償で協力を受けた．

2.3 タスク支援ソフトウェア：市民参加による固有表現抽出システム概要

市民参加による固有表現抽出システムと Linked Data 自動変換システム（図 1）との関係を**図 2** に示す．本抽出システムでは，フロントエンドのウェブアプリケーションであるタスク支援ソフトウェアを，バックエンドのシステム管理機能の下で稼働させる構成をとっている．タスク参加者は，ウェブブラウザ上で以下にあげる操作を行う．

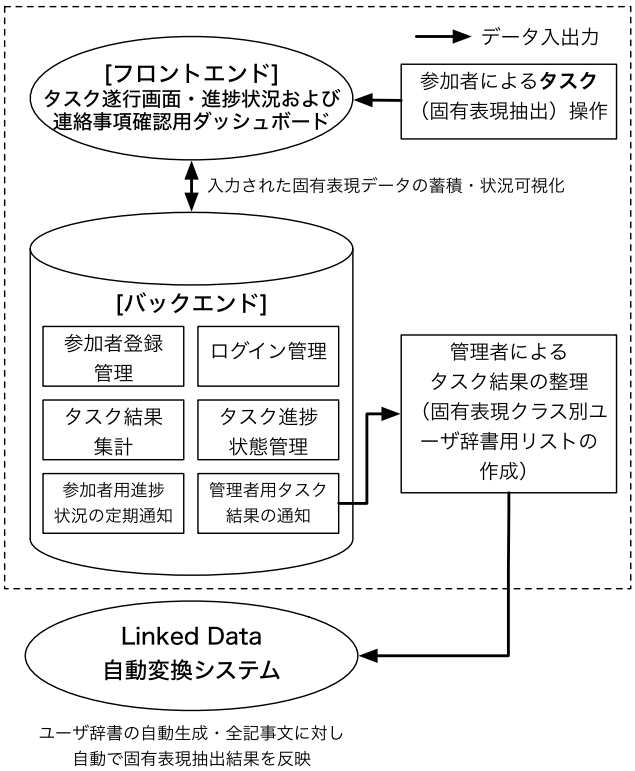


図 2 市民参加による固有表現抽出システム（点線枠内）の構成および Linked Data 自動変換システムとの関係（文献 [8] より引用）

Fig. 2 Configuration of a named entity extraction system with citizen participation (in the dotted box) and its relationship with the automated Linked Data conversion system.

まず，タスク参加者および後述の管理者はログイン後，ダッシュボードと呼ぶ固有表現抽出済みの記事文総数および月別総数や個人のタスク完了記事文数などを確認できる画面で進捗状況を把握する．タスクを行う場合は，タスク遂行画面に移動する．その際，バックエンドでは，タスク参加者のログイン日時，操作内容，固有表現抽出の進捗を記録し，データベースに蓄積する．

実際のタスク遂行画面では，未解析の目録記事文の 1 つがバックエンドで決定され，ランダムかつ他のタスク参加者と重複を避けて表示される[*3]．

次に，表示した記事文に対して，あらかじめ形態素解析ツール MeCab [18] とそのユーザ辞書を用いて自動的に固有表現抽出を行う．その抽出結果をタスク参加者が確認して，修正点があれば正しい単語と対応する固有表現ラベルを登録する．修正点がなければ，そのまま修正完了としてタスクを終了し，新しい記事文を表示する．なお，タスク参加者の抽出結果を見直したい場合は，再度修正できる機能も付加している．

なお，本ソフトウェアを使用する前に，タスク参加者に対して操作説明会を開いて画面操作や固有表現抽出に必要なルールの具体例などを共有し，後に参加者が単独でタスクを行う際の疑問点や不安を解消した．

[*3] 具体的な操作画面は文献 [8] の図 3-1，図 3-2，図 3-3 で参照できる．

#	固有表現種類名	単語（固有表現）	単語数
42	人名 Person	肥州様	1000
6173	地名 Place	肥州	6
8399	人名 Person	窪田肥州	3
9284	人名 Person	肥州	3
10627	人名 Person	松平肥州	2
14951	人名 Person	松浦肥州	1

図 3 固有表現リスト [19] で「肥州」を検索した結果．各単語に付与されたリンク先は小城藩日記データベースでの検索結果となる．

Fig. 3 Results of a search for “Hishu (肥州)” in the list of named entities [19]. The link given to each word is the search result in the Ogi-han Nikki Database.

2.4 タスク参加者による抽出結果の質の担保

タスク参加者には一定の候文読解能力があるが，クラス分け抽出ルールの下でも，当然ながら解釈の仕方に個人差あるいは誤りが生じる．

そこで，本研究では，タスク参加者の抽出結果を直接正解データとして用いず，バックエンドから出力されたクラス別固有表現データを随時整理する管理者を設置した．管理者は，抽出された固有表現が妥当なクラスに仕分けられているかを目視で確認する．

その後，確認済み固有表現データを Linked Data 自動変換システムに投入することで，新しい辞書が自動的に再構築される．この再構築した辞書を用いて形態素解析および固有表現抽出が実行されると，全目録記事文の固有表現抽出結果も自動的にアップデートされる．

加えて，タスク参加者および管理者は，最新の抽出状況を参照できるウェブサイト「固有表現リスト」[19] で，固有表現がどのクラスに何回判定されたかを参考にできる．たとえば，「肥州」という語に対し，クラス判定に迷う場合，固有表現リストのサイトで「肥州」を検索すると，これまでに人名および地名と判定されており，人名とされている回数が地名より多いことを参照できる（**図 3**）．

以上の仕組みにより，タスク参加者が抽出した固有表現データの質を担保した．

3. ディープラーニングによる江戸期候文からの固有表現抽出手法

3.1 課題

従来の Linked Data 自動変換システムおよび前章の市民参加型固有表現抽出手法では，形態素解析ツールのユーザ辞書内容の充実により，固有表現抽出精度を上げる手法をとった．これらの手法は，記事文の蓄積が少なかった初期段階において，ユーザ辞書の登録内容がほぼ全記事文に対して適用されるため有効であった．しかし，記事文が 2，3 万件を超え，ユーザ辞書がアップデートされるにつれて，以下にあげる新たな課題が生じた．

- 形態素解析ツールのユーザ辞書に登録されていない，あるいは既出の記事文にいっさい出現しない未知語は，自動的に固有表現抽出できない．
- タスク参加者間で固有表現クラスの判定に多少の相違がある．または，多義語のために判定基準に曖昧さが生じることがある．そのため，管理者による固有表現データの確認が手間どるようになった．

これらの課題は，形態素解析ツールのユーザ辞書に固有表現を登録する手法では，大量の未知語に対処することに限界があることを示唆している．

3.2 既往研究

本研究では，2 章で述べた候文中で使用されている地域特有の固有表現の収集に加え，未知の候文に対して固有表現とそのクラスを高精度に推測可能にする 1 つの方法として，文中での個々の固有表現が文脈に応じてどのように使われているかをディープラーニングで学習することを試みた．

一般に，固有表現抽出のような自然言語処理タスクは，時系列データの予測問題として取り扱われている．ある文で特定の単語が出現する確率を求めるために最も重要な情報は，単語間の類似度である．

従来，単語間の類似度を求めるには，WordNet [20] など人手で作られたシソーラスや単語どうしの関係を形式的に記述可能なオントロジ [21] を利用するのが主流であったが，英語や日本語など生きた言語のシソーラスを手作業で構築し続けることは現実的ではない．

そこで，「同じ文脈で出てくる言葉は似たような意味を持つ傾向がある」と考える分布仮説 [22] に基づき，単語分散表現（Word embeddings）[23] と呼ばれる，単語間の類似度を高次元の実数ベクトルで数値化して表現する手法が開発された．ここでの文脈とは，コーパス内の文を構成する各単語の周辺にある単語群のことを指す．

単語分散表現を獲得するツールとしては，単語分散表現の学習ツールである Word2vec [24] に含まれる Skip-gram（Continuous Skip-Gram Model）[24] および CBOW（Continuous Bag-of-Words Model）[24] のほか，GloVe [25], fastText [26] などがあげられる．

しかし，上記の単語分散表現は，文脈を考慮しておらず，多義語ではいくつかのベクトル値が 1 つにまとまった値として算出されるため，江戸期の記録に頻出する多義語，たとえば場所としての「肥州」と役職名としての「肥州」のような語には適切に対応できない．

多義語に対応するには，2018 年以降の ELMo（Embeddings from Language Models）[27]，BERT（Bidirectional Encoder Representations from Transformers）[10]，Flair [11] など，同じ表記であっても文脈に応じて異なるベクトル値を保持できる単語分散表現の構築手法を採用する必要がある．

これらの手法は，いずれも Transformer [28] における Attention 機構 [29] を活用した手法である．Attention とは，入力中の単語どうしに加え，Wikipedia 記事などから構築された汎用的かつ大規模な単語分散表現である事前学習データを利用して，各単語と自身以外の単語との関連度（注目度）の高さを動的に算出する手法である．この事前学習データのベクトル値に対し，ドメインに特化した教師データを用いて補正をかける追学習（Fine-tuning）を行えば，大量に教師データとなる例文を集めることが困難であっても高精度な処理結果を得られることになる．

3.3 提案手法

我々は，抽出対象の文が候文であっても，候文中に出現する単語や文字およびそれらの意味が現代日本語と大きくかけ離れていなければ，上記 Attention を用いる手法を候文に対して応用可能であると考えた．なぜなら，比較的現代に近い江戸期の候文は，漢語を多く含んでいるものの，日記目録における候文翻刻では，旧字体を新字体に統一してデジタルアーカイブでの検索を簡便化していることに加えて，MeCab による形態素解析用基本辞書として現代日本語用の Unidic [30], [31]，ユーザ辞書として候文固有の単語がそれぞれ登録されている状態で，登録済みの固有表現であれば，実際に候文から抽出できているためである．

また，上記の事前学習データと追学習用データを組み合わせ，候文に対する固有表現判定モデルを生成する際に，再現性および可用性を高めるため，本稿では，日本語を含めて様々な単語分散表現データを利用でき，固有表現判定モデル構築および評価に必要なプログラムをまとめて扱える Flair フレームワーク [32] およびそのライブラリ [33] を採用した．

さらに，本研究の目的は，文中に含まれる固有表現の位置と長さ，クラス分けを推測できるモデルの構築であり，あらかじめ形態素解析による単語切り出し済みの事前学習モデルの使用は適切でない．**表 2** に示すとおり，候文 73,984 文中の全文字数は 1,146,912 で，そのうち 95.07%（1,090,357）

表 2 「小城藩日記データベース」に収録されている目録記事文 73,984 件中の文字種別カウント数および割合

Table 2 Counts and percentages of 73,984 articles in "Ogi-han Nikki Database" by character type.

	カウント数	全文字中で占める割合 (%)
全文字	1,146,912	100.00
漢字	1,090,357	95.07
ひらがな	29,097	2.54
カタカナ	20,992	1.83
その他文字種	6,466	0.56

城	B-JINMEI	池	B-JINMEI
州	I-JINMEI	田	I-JINMEI
様	O	与	I-JINMEI
本	B-PLACE	四	I-JINMEI
行	I-PLACE	右	I-JINMEI
寺	I-PLACE	衛	I-JINMEI
へ	O	門	I-JINMEI
御	O	大	B-PLACE
葬	B-EVENT	坂	I-PLACE
礼	I-EVENT	と	B-TERMS
之	B-TERMS	参	B-EVENT
事	I-TERMS	着	I-EVENT
		之	B-TERMS
		事	I-TERMS

図 4 IOB2 タグ形式データの例（文献 [9] より引用）．JINMEI は PERSON（人名クラス）の意

Fig. 4 An example of IOB2 format. JINMEI means PERSON class.

が漢字を占め，1 つの漢字で意味をなす単語も多い．そのため，形態素解析の結果が反映されない文字レベルで計算された Flair Embeddings [34] と呼ぶ独自方式の日本語事前学習データおよび**図 4** に示す IOB2 タグ形式 [35] でのラベリング手法で作成した追学習データを用いることとした．

本研究の事前学習に用いた日本語 Flair Embeddings は，Wikipedia 日本語版ダンプファイル（2018/12/20）から生成されたものである [36]．

4. 固有表現抽出手法の評価

4.1 市民参加によるクラス別固有表現抽出評価結果の推移

4.1.1 概要

本研究では，市民参加による固有表現抽出の確認作業を 2019 年 5 月から 2020 年 5 月まで行った．この間，**図 5** および**表 3** に示すとおり，日記目録記事文が 2019 年 5 月時点の 18,317 件から同年 6 月，9 月，2020 年 3 月にそれぞれ 29,221 件，41,719 件，50,801 件が随時データベースに追加され，最終的には 2020 年 10 月に全目録記事 73,984 件の翻刻とデータベース登録が完了した．本作業の範囲は，

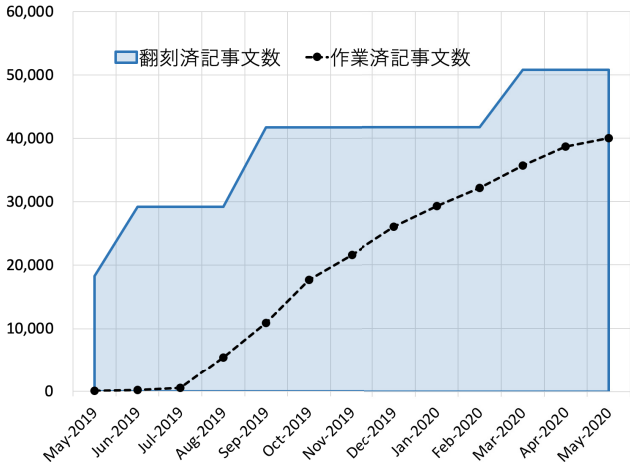


図 5 市民参加期間における翻刻済記事文数およびタスク済記事文数

Fig. 5 Number of reprinted articles and content verified during the public participation period.

表 3 目録記事文増補期間における固有表現の抽出集計月，総記事文数および seqeval での検証時に除外した記事文数

Table 3 Work aggregation month, total number of articles, and number of articles excluded during validation with seqeval in period of the named-entity extraction and expanding number of the catalog (Mokuroku) articles.

抽出集計月	総記事文数	除外記事文数
May-2019	18,317	1,786
Jun-2019	29,221	125
Jul-2019	29,221	125
Aug-2019	29,221	125
Sep-2019	41,719	9
Oct-2019	41,719	9
Nov-2019	41,719	9
Dec-2019	41,719	9
Jan-2020	41,719	9
Feb-2020	41,719	9
Mar-2020	50,801	2
Apr-2020	50,801	2
May-2020	50,801	2
Jun-2020	70,289	2
Jul-2020	70,289	2
Aug-2020	72,739	0
Sep-2020	72,739	0
Oct-2020	73,984	0

目録記事文に割り振られた一意の登録番号の 1 から 40,000 までを対象とした．

また，市民が参加した 2019 年 7 月以降，固有表現抽出クラス分けルールについて確認を数度行うことで抽出効率の向上を図った．その結果，抽出の目標としていた記事文数 40,000 件（全記事文の 54.1%）の確認をおよそ 12 カ月で完了した（図 5）．

管理者による MeCab ユーザ辞書の更新は翻刻文のデータベース登録や新規辞書登録が貯まった時点で不定期にま

とめて行い，全タスク完了後の 2020 年 10 月時点で更新を終了した．

抽出精度の測定については，毎月 1 回，Linked Data 自動変換システムを稼働させることで，各時点の登録済記事文に対する固有表現抽出の状態をデータとして保存した．そして，全 40,000 件の管理者による内容確認済みの状態を基準（正解データ）とし，保存した過去の固有表現抽出結果と比較することで抽出精度の指標である適合率（Precision），再現率（Recall）および F 値（F-measure）を測定した．具体的には，本測定での適合率とは，Linked Data 自動生成システムによる固有表現抽出結果の中でクラス判定まで正解だったものの割合である．再現率は，全正解データに対して，同システムの固有表現抽出結果およびそのクラス判定まで正解だったものの割合である．いい換えると，適合率は固有表現抽出の正確さ，再現率は固有表現として認識できる感度の高さを表している．F 値とは，上記定義で算出した適合率と再現率の調和平均である．

測定には，系列ラベリング問題関連の精度判定専用 Python ライブラリである seqeval [37] を使用した．ただし，seqeval では，テストする文と正解データとする文が完全一致する必要がある．実際にデータベースに登録されている記事文の中には，過去に異体字や誤りなどを含んでいたため，気がついた時点で修正されたものがある．そのため，正解データと異なる記事文は測定から除外している．登録記事文から除外した実際の数は，表 3 に示す．

4.1.2 候文の固有表現抽出での評価基準

なお，候文の固有表現抽出では，固有表現自体をあらかじめ認識する必要があり，そのうえでクラスを文脈で判断しなければならない．そのため，本研究では，文中の固有表現自体を見逃さないことを，クラスを正しく判定することよりも重視している．すなわち，適合率より再現率を重んじている．さらに，F 値が 0.9 以上である場合は，文脈から固有表現を明確に認識できているととらえ，「高精度」と評価する．

4.1.3 結果

まず，市民参加によるタスク実行期間内での MeCab ユーザ辞書のデータ追加および整理作業の結果として，2020 年 10 月時点での抽出済み固有表現の出現数，出現数の割合および固有表現数を**表 4** に示す．この表での固有表現の出現数とは，文中での固有表現の出現回数をカウントしたものである．出現数の割合とは，全固有表現の出現数の中で各クラスが占める割合である．

また，各クラスにおける固有表現を出現度数順に並べた際の出現度数から算出される占有率（分布）を**図 6**，**図 7**，**図 8**，**図 9**，**図 10**，**図 11**，**図 12** に示した．たとえば，人名クラス（図 6）の場合，固有表現出現数の上位 100 位までででこのクラス全体の 26.4%を占め，1,000 位までで 65.9%を占有することを示す．これに対し，候文用語クラス（図 7）では，上位 1 位の辞書登録でこのクラスの全出現数の 30.1%，上位 100 位までの登録では 98.8%をそれぞれ占有する．

表 4 市民参加型タスク終了時（2020 年 10 月時点）の記事文 73,984 件に出現した固有表現の出現数，全固有表現の出現数に対する各クラスの占める割合，および固有表現数

Table 4 The number of occurrences and types of each named entity class in 73,984 titles at the end of the citizen participation task (as of October 2020), the ratio of each class in the total number of named entities' occurrences, and the number of named entities.

クラス名	固有表現の出現数	出現数の割合（%）	固有表現数
全クラス	409,004	100.0	18,141
出来事	131,032	32.0	5,781
候文用語	117,384	28.7	257
人名	52,098	12.7	7,238
役職・役割	50,607	12.4	2,019
場所	37,685	9.2	1,897
数量	10,931	2.7	809
日時	9,267	2.3	581

さらに，各クラスにおける占有率は，頻出の固有表現の個数が上昇するにつれて単調増加することから，未知語の登録は正確な固有表現のクラス判定に寄与する．特に，人名クラスおよび出来事クラスにおいては，固有表現が多様であることから，固有表現の抽出とクラス判定の精度向上には多数の未知語への対応が必要であった．

次に，各クラスごとの適合率，再現率および F 値を**図 13**，**図 14**，**図 15**，**図 16**，**図 17**，**図 18**，**図 19** に示す．

なお，ユーザ辞書は，2019 年 6 月，同年 7 月，同年 9 月，2020 年 2 月，同年 6 月に一定数新しい固有表現が蓄積した時点で管理者が整理し，Linked Data 自動変換システムにデータを投入して再構築した．

続いて，図 2 のバックエンド「タスク進捗状態管理」で蓄積した各参加者のタスク結果を，Linked Data 自動変換システムに蓄積されている最終確認済みかつ最新の固有表現データと比較し，正解率を求めた（**図 20** および**表 5**）．ただし，正解率を求める際は，参加者が誤りと判断して能動的に修正した（修正データが存在する）記事文を対象にしており，修正の必要がなくそのまま完了としたものは修正データがないため除外した．また，全参加者 8 名中 1 名は，修正データのある記事文がなかったため，除外した．

4.1.4 考察

タスク実行期間中，新しい参加者が追加された 2019 年 5 月と同年 7 月および一定の固有表現データが蓄積した 2020 年 2 月に，抽出精度の一時的な低下が見られた（図 13 から図 19）．

その際，2.4 節で示したとおり，管理者がタスク結果を整理し，固有表現抽出ルールを参加者と複数回見直すことで，個人間の生じた判定の差異を修正した．加えて，参加者が判定した結果を参照できるように，これまで蓄積した

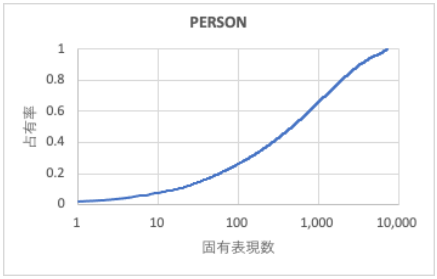


図 6 人名（PERSON）クラスの固有表現数に対する占有率
Fig. 6 Occupancy rate of PERSON class relative to its number of named entities.

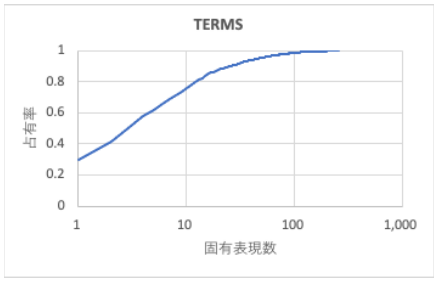


図 7 候文用語（TERMS）クラスの固有表現数に対する占有率
Fig. 7 Occupancy rate of TERMS class relative to its number of named entities.

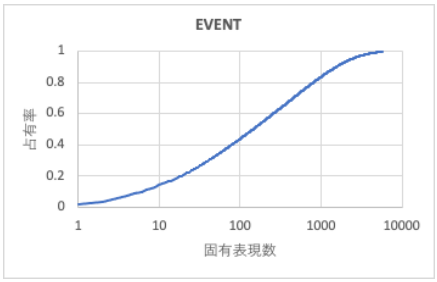


図 8 出来事（EVENT）クラスの固有表現数に対する占有率
Fig. 8 Occupancy rate of EVENT class relative to its number of named entities.

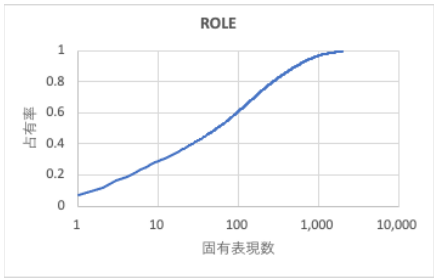


図 9 役職・役割（ROLE）クラスの固有表現数に対する占有率
Fig. 9 Occupancy rate of ROLE class relative to its number of named entities.

固有表現リストを検索できる Web サイトを設置した [19]．このような措置と管理者による固有表現データの整理で精度は再度向上した．

固有表現リストの蓄積と整理を進めるなか，クラス間で抽出難易度には差があることが明らかになった．

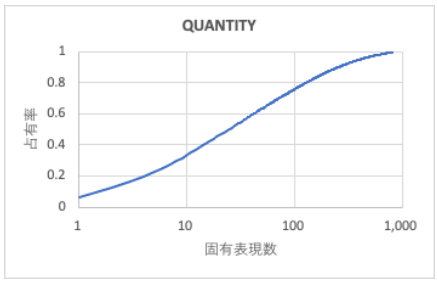


図 10 数量（QUANTITY）クラスの固有表現数に対する占有率
Fig. 10 Occupancy rate of QUANTITY class relative to its number of named entities.

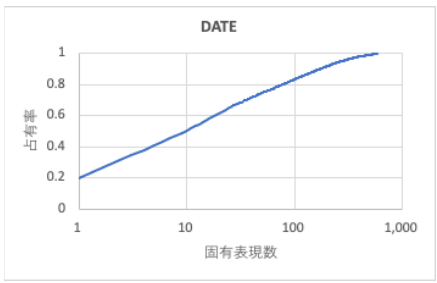


図 11 日時（DATE）クラスの固有表現数に対する占有率
Fig. 11 Occupancy rate of DATE class relative to its number of named entities.

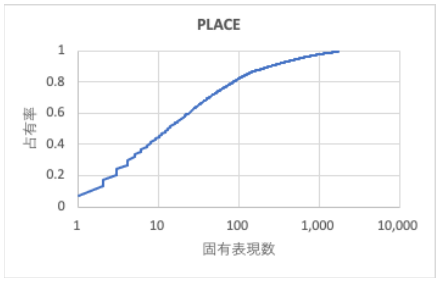


図 12 場所（PLACE）クラスの固有表現数に対する占有率
Fig. 12 Occupancy rate of PLACE class relative to its number of named entities.

まず，人名と候文用語の各クラスでの抽出精度は他のクラスよりつねに高かった（図 13 および図 14）．図 20 および表 5 の参加者別タスク正解率においても，他のクラスに比べて正解率が高く，参加者間の差が小さかった．

次に，全記事文中に占める固有表現の出現数が 3 割を超える出来事クラスについては，2019 年 9 月以降に再現率が適合率を上回り，参加者の正解率（図 20）では人名や候文用語の次に高い結果を示した．

役職・役割，数量，日時および場所については，これらの順で正解率が低下した（図 20）．

一般に，これらのクラスに属する語は多義的であり，文脈から判定せざるをえない場合が多く，所定のクラス判定のルールのみでは判断が難しい．また，「御」のような冠詞や「様」などの接尾辞を含むか含まないか，複数の単語が連なった固有表現をどこまでまとめるかなど，文中のどの

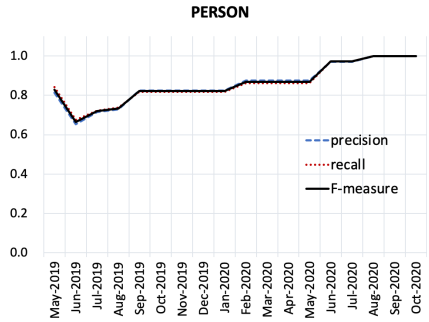


図 13 人名（PERSON）クラスの F 値，適合率および再現率
Fig. 13 F-measure, precision and recall of the PERSON class.

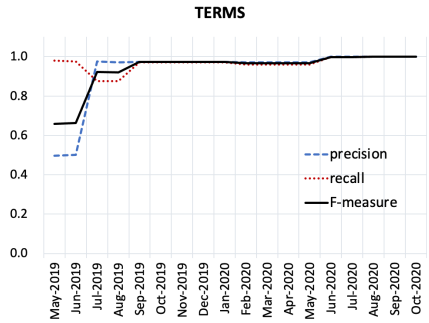


図 14 候文用語（TERMS）クラスの F 値，適合率および再現率
Fig. 14 F-measure, precision and recall of the TERMS class.

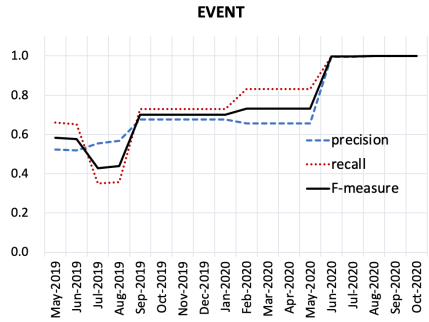


図 15 出来事（EVENT）クラスの F 値，適合率および再現率
Fig. 15 F-measure, precision and recall of the EVENT class.

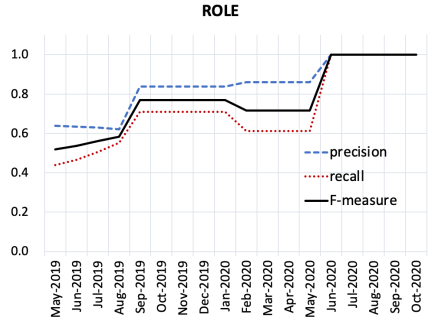


図 16 役職・役割（ROLE）クラスの F 値，適合率および再現率
Fig. 16 F-measure, precision and recall of the ROLE class.

部分を固有表現として抽出するか判断が難しい語が多く含まれている．そのため，クラスの誤判定が多いことが示唆される．

その一方，これらのクラスは適合率が再現率を上回って

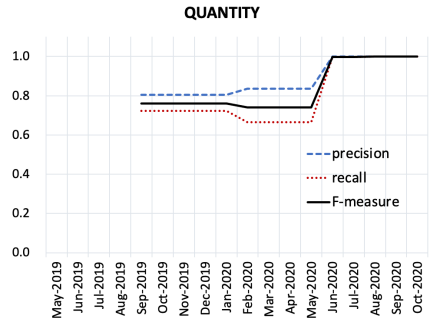


図 17 数量（QUANTITY）クラスの F 値，適合率および再現率．2019 年 8 月から抽出開始
Fig. 17 F-measure, precision and recall of the QUANTITY class. In August of 2019, the extraction began.

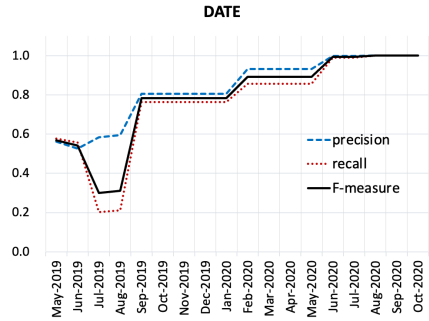


図 18 日時（DATE）クラスの F 値，適合率および再現率
Fig. 18 F-measure, precision and recall of the DATE class.

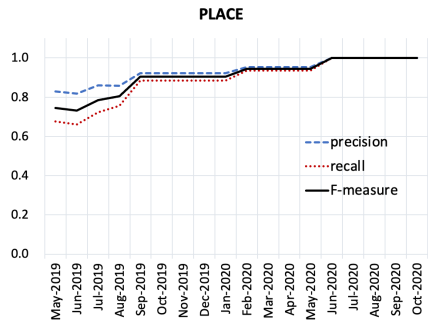


図 19 場所（PLACE）クラスの F 値，適合率および再現率
Fig. 19 F-measure, precision and recall of the PLACE class.

いた（図 16，図 17，図 18 および図 19）．これは，クラス判定は不正解だったものの固有表現の抽出そのものはできている場合が多いことを示している．さらに，抽出すべき固有表現の種類数自体は，役職・役割，数量，日時および場所においてそれぞれ 2,019，809，581 および 1,897 であり（表 4），人手での抽出は量的にも困難ではない．

したがって，クラス判定ルールをより明確にして辞書による再利用効果を高める，あるいは個別での対処に注力するなどの工夫で抽出精度をさらに向上させる余地があると考えられる．

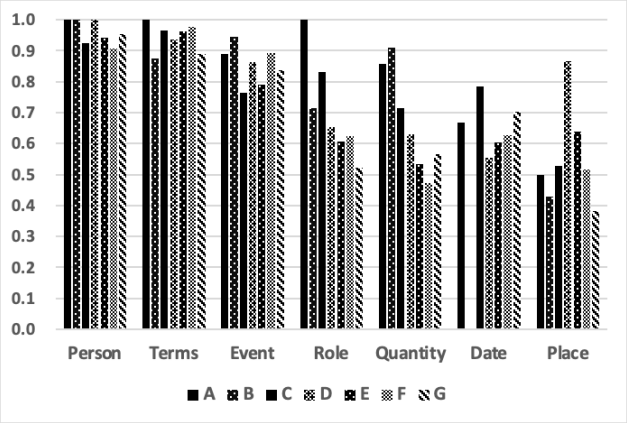


図 20 固有表現クラスおよび参加者別タスク正解率．A から G までは参加者で A および B は無償，C から G までは有償での参加者である

Fig. 20 Named entity class and task correctness by participant; A to G are participants, A and B are free, and C to G are paid participants.

表 5 参加者別タスク正解率とその平均および参加者 A から G が能動的に修正した固有表現の総数（修正数）

Table 5 Percentages of task correctness actively modified by participants and their means, and total number of named entity extraction by participants A to G (number of modified named entities).

参加者	A	B	C	D	E	F	G	全体平均
人名	1.0000	1.0000	0.9256	1.0000	0.9418	0.9074	0.9534	0.9612
候文用語	1.0000	0.8750	0.9667	0.9355	0.9614	0.9786	0.8889	0.9437
出来事	0.8889	0.9459	0.7636	0.8630	0.7900	0.8927	0.8358	0.8543
役職・役割	1.0000	0.7143	0.8306	0.6538	0.6074	0.6256	0.5225	0.7077
数量	0.8571	0.9091	0.7144	0.6296	0.5339	0.4724	0.5658	0.6689
日時	0.6667	n/a	0.7835	0.5556	0.6032	0.6266	0.7043	0.6566
場所	0.5000	0.4286	0.5271	0.8667	0.6388	0.5157	0.3834	0.5515
平均	0.8447	0.8121	0.7874	0.7863	0.7252	0.7170	0.6934	0.7656
修正数	85	124	13,661	170	5,738	12,917	7,283	5,711

4.2 ディープラーニングによるクラス別固有表現抽出精度の検証

4.2.1 概要

3 章で示した Flair フレームワークによる固有表現抽出および精度判定を行った．具体的な手法は，文献 [9] に示す．

追学習および精度判定に用いる文は，目録記事文の登録番号 1 から 40,000 までの文からランダムに 5,000 件刻みで必要件数を選択したものである（以降，5,000 件から 40,000 件までの記事文の各集合をデータセットと呼ぶ）．

さらに，Flair 組み込み機能は，各データセットを追学習用および精度判定用サブセットに 9 対 1 の割合で自動分割する．各サブセットは互いに独立し，追学習用とされたサブセットが精度判定に使用されることはない．

また，F 値などの抽出精度の測定は，事前学習と追学習（Fine-tuning）で固有表現判定モデルを構築する際に行った．実行した判定モデル構築およびそのモデルを用いた固有表現抽出プログラムは，Google Colab [38], [39] で参照可能である．

表 6 ディープラーニングで適合率，再現率および F 値が 0.9 以上になるために必要な追学習用記事文数

Table 6 The number of fine-tuning articles required for deep learning to achieve a precision rate, a recall rate, and an F-measure of 0.9 or higher.

	適合率	再現率	F 値
TERMS	5,000	5,000	5,000
PERSON	5,000	5,000	5,000
PLACE	15,000	15,000	15,000
EVENT	20,000	15,000	20,000
ROLE	25,000	20,000	25,000
DATE	40,000	40,000	40,000
QUANTITY	なし	30,000	40,000

毎回ランダムに作成された各データセットで 3 回ずつ判定モデルを構築して，適合率，再現率および F 値の測定を行った．

4.2.2 結果および考察

図 21，**図 22**，**図 23**，**図 24**，**図 25**，**図 26**，**図 27** に，データセット量が変化した際の固有表現クラス別 F 値の平均，最小値および最大値を示す．

その結果，データセットが 5,000 件から 40,000 件に増加するにともない，全クラスの F 値が向上し，その最小値と最大値の差が小さくなった．また，0.9 以上の F 値を超えるために必要な記事文数は，**表 6** に示すとおり，クラスによって差異が見られた．

具体的には，人名および候文用語クラスは，5,000 件の学習で F 値 0.9 以上を示した（表 6）．これらのクラスの次に，場所，出来事，役職・役名の順で，20,000 件程度の学習により高精度な結果を得られた．対して，日時および数量クラスでは，追学習データを記事文の半数以上である 40,000 件を投入することで F 値が 0.9 を超えた．

最終的に，2020 年 10 月時点までに記事文から抽出した全固有表現数 409,004 に各クラスが占める割合を考慮すると，25,000 件（全記事文の約 33.8%）の正解データセットがあれば，表 4 で示した全固有表現の数で 95.0%を占める人名，候文用語，場所，出来事，役職・役割の各クラスで，F 値 0.9 を超えると示唆される．

これらの結果から，記事文の文脈情報を含んだ学習データセットから得られる単語分散表現の利用は，従来の形態素解析用辞書を基にした固有表現抽出手法の課題を解決する手がかりとなることが明らかになった．

5. まとめ

わが国には，小城藩日記のような候文で書かれた江戸期約 260 藩の業務記録，証文などの古記録・公文書が膨大に残存する．地域の詳細な歴史が記述されている貴重な文書

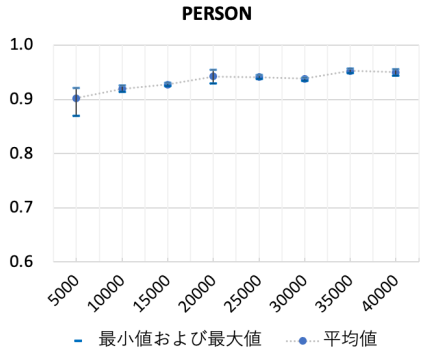


図 21 人名クラスにおける各データセット学習時の F 値平均
Fig. 21 Average of F-measure for each dataset during training in PERSON class.

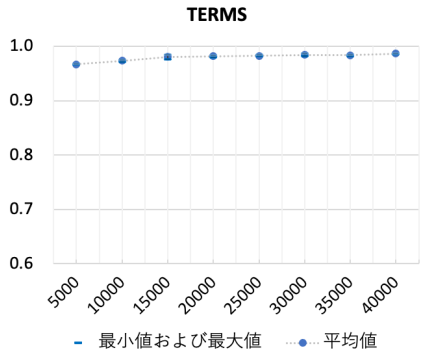


図 22 候文用語クラスにおける各データセット学習時の F 値平均
Fig. 22 Average of F-measure for each dataset during training in TERMS class.

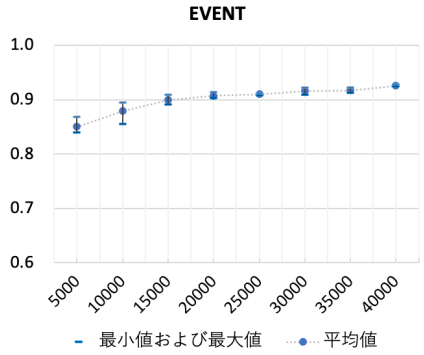


図 23 出来事クラスにおける各データセット学習時の F 値平均
Fig. 23 Average of F-measure for each dataset during training in EVENT class.

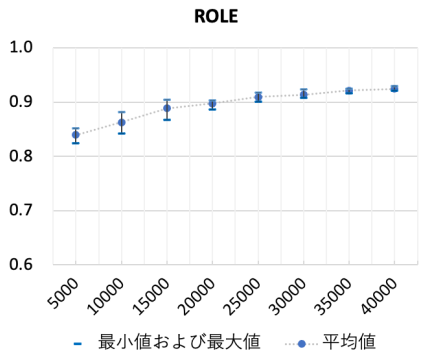


図 24 役職・役割クラスにおける各データセット学習時の F 値平均
Fig. 24 Average of F-measure for each dataset during training in ROLE class.

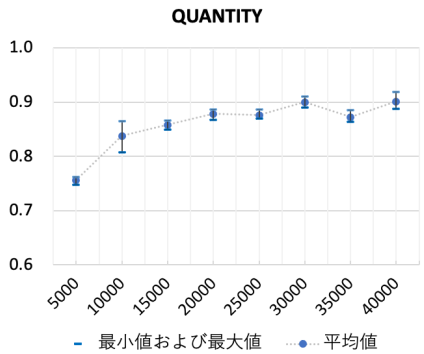


図 25 数量クラスにおける各データセット学習時の F 値平均
Fig. 25 Average of F-measure for each dataset during training in QUANTITY class.

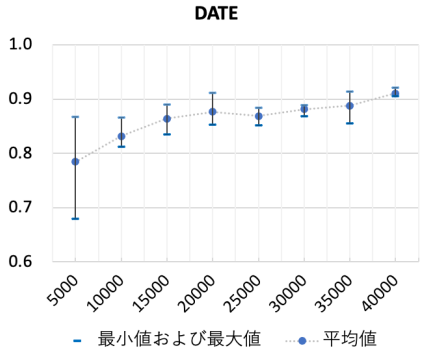


図 26 日時クラスにおける各データセット学習時の F 値平均
Fig. 26 Average of F-measure for each dataset during training in DATE class.

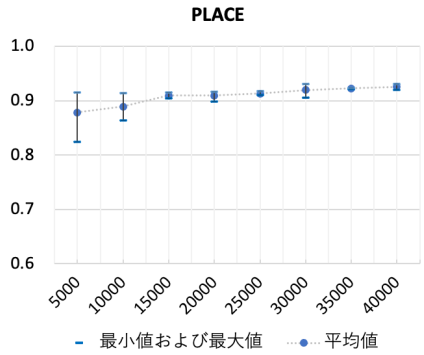


図 27 地名クラスにおける各データセット学習時の F 値平均
Fig. 27 Average of F-measure for each dataset during training in PLACE class.

を，災害や経済的理由などで消失する前に，Web で多くの人が参照し，必要に応じて分析可能なデータとして機械可読化しておくことは，今後ますます重要となる．

そのためには，データベース構築に，コンピュータビジョン，自然言語処理などの最新技術をうまく取り入れることに加え，地域市民の協力を得て，分析可能なデータの在り方をともに考え，効率良く作り上げる場にすることが必要である．

本研究では，小城藩日記データベースを事例に，江戸期に候文で書かれた地域の古記録を，可能な限り人的コストを抑えて機械可読化していく手法を提案した．

具体的には，Web で収集困難な地域特有の固有表現を，候文の読み解きや地元の固有名詞を識別できる市民が収集し，最終的に固有表現クラスのラベルがついた教師（正解）データを作成した．ただし，このような市民参加による手法は，一定の専門知識を持つ参加者による確認ができる一方で，固有表現の判定に個人間あるいはクラス間でばらつきが生じ，データのさらなる整理が必要となるため限界がある．

そこで，江戸期候文の教師データと現代日本語の単語分散表現データである事前学習を組み合わせ，多義語を解釈可能な固有表現判定モデルを構築した．その結果，市民参加による固有表現抽出手法のみでは収集に限界のあった未知かつ地域特有の固有表現を，判定モデルから高精度に抽出可能であることが明らかになった．すなわち，実在するデータセットを用いて地域資料の機械可読化を実践的に示すことで，他の古記録の Linked Data 化を含む機械可読化が一層容易なものとなる可能性を示した．

今後は，小城藩日記データベースの目録文に紐づく日記本文やそれ以外の江戸期古記録に対して検証と技術開発を進め，機械可読化への汎用性を高めていく予定である．

謝辞 本研究は，JSPS 科研費 JP19K20630 の助成を受けたものである．また，2 章で述べたタスクの遂行にあたり，多大な貢献をいただいた市民の方々に感謝の意を表する．

参考文献

[1] The TEI Consortium: Text Encoding Initiative, TEI: Text Encoding Initiative (online), available from ⟨https://tei-c.org⟩ (accessed 2021-04-25).

[2] Heath, T. and Bizer, C.: *Linked data: Evolving the web into a global data space*, Morgan & Claypool Publishers (2011).

[3] The National Museum of Japanese History, Goto, M., Nakamura, S., Nishioka, C., Puspita, A.A., Yamada, T., Hashimoto, Y., Yoshiga, N., Sekino, T., Kokaze, N. and Yamasaki, S.: *Japanese and Asian Historical Research in the Digital Age*, The National Museum of Japanese History (2021).

[4] 佐賀大学地域学歴史文化研究センター：小城藩日記データベース，ホームページ（オンライン），入手先 ⟨https://crch.dl.saga-u.ac.jp/nikki/⟩（参照 2021-04-17）．

[5] 吉賀夏子, 只木進一, 伊藤昭弘：小城藩日記データベースの構築，研究報告人文科学とコンピュータ（CH），Vol.2018-CH-117, No.3, pp.1–7 (2018)（オンライン），入手先 ⟨http://id.nii.ac.jp/1001/00187419/⟩．

[6] 吉賀夏子，只木進一：古典籍書誌データ構造に対応した Linked Data への半自動変換，情報処理学会論文誌，Vol.59, No.2, pp.257–266 (2018).

[7] Grishman, R. and Sundheim, B.: Message Understanding Conference-6: A Brief History, *Proc. 16th Conference on Computational Linguistics – Volume 1*, *COLING '96*, pp.466–471, Association for Computational Linguistics (online), DOI: 10.3115/992628.992709 (1996).

[8] 吉賀夏子，只木進一：低コストな Linked Data 化を目指したクラウドソーシングによる固有表現収集の試み，じんもんこん 2019 論文集，Vol.2019, pp.239–244 (2019).

[9] 吉賀夏子，堀 良彰，永崎研宣：候文における文字単位の単語分散表現モデルに基づく固有表現抽出手法，研究報告人文科学とコンピュータ（CH），Vol.2021-CH-125, No.4, pp.1–7 (2021)（オンライン），入手先 ⟨http://id.nii.ac.jp/1001/00209267/⟩．

[10] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[11] Akbik, A., Blythe, D. and Vollgraf, R.: Contextual string embeddings for sequence labeling, *Proc. 27th International Conference on Computational Linguistics*, pp.1638–1649 (2018).

[12] Vohland, K., Land-zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R. and Wagenknecht, K. (Eds.): *The Science of Citizen Science*, Springer Nature (2021).

[13] Tauginienė, L., Butkevičienė, E., Vohland, K., Heinisch, B., Daskolia, M., Suškevičs, M., Portela, M., Balázs, B. and Prūse, B.: Citizen science in the social sciences and humanities: The power of interdisciplinarity, *Palgrave Communications*, Vol.6, No.1, p.89 (online), DOI: 10.1057/s41599-020-0471-y (2020).

[14] 国立歴史民俗博物館，東京大学地震研究所，京都大学古地震研究会：みんなで翻刻，ホームページ（オンライン），入手先 ⟨https://honkoku.org⟩（参照 2021-04-21）．

[15] Library of Congress: BY THE PEOPLE, homepage (online), available from ⟨https://crowd.loc.gov⟩ (accessed 2021-04-21).

[16] 関西大学アジア・オープン・リサーチセンター（KU-ORCAS）：コロナアーカイブ@関西大学，関西大学コロナアーカイブ（オンライン），入手先 ⟨https://www.annex.ku-orcas.kansai-u.ac.jp/s/covid19archive/page/covidmemory⟩（参照 2021-04-21）．

[17] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B. and Allahbakhsh, M.: Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions, *ACM Computing Surveys* (*CSUR*), Vol.51, No.1, pp.1–40 (2018).

[18] Kudo, T.: MeCab: Yet Another Part-of-Speech and Morphological Analyzer (ver. 0.996), homepage (online), available from ⟨http://taku910.github.io/mecab/⟩ (accessed 2021-04-17).

[19] 吉賀夏子：小城藩日記データベースの固有表現リスト，小城藩日記プロジェクト（オンライン），入手先 ⟨https://winter.ai.is.saga-u.ac.jp/cs/ne-words.php⟩（参照 2021-04-25）．

[20] Miller, G.A.: WordNet: A Lexical Database for English, *Comm. ACM*, Vol.38, No.11, pp.39–41 (online), DOI: 10.1145/219717.219748 (1995).

[21] Gruber, T.R.: *Ontology* (*Entry in the Encyclopedia of Database Systems*), Springer-Verlag (2008).

[22] Harris, Z.S.: Distributional structure, *Word*, Vol.10, No.2-3, pp.146–162 (1954).

[23] Turian, J., Ratinov, L.-A. and Bengio, Y.: Word Representations: A Simple and General Method for Semi-Supervised Learning, *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, pp.384–394, Association for Computational Linguistics (2010) (online), available from ⟨https://www.aclweb.org/anthology/P10-1040⟩.

[24] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013).

[25] Pennington, J., Socher, R. and Manning, C.D.: Glove:

Global Vectors for Word Representation, *EMNLP*, Vol.14, pp.1532–1543 (2014).
[26] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association for Computational Linguistics*, Vol.5, pp.135–146 (2017).
[27] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep contextualized word representations, *Proc. NAACL* (2018).
[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I.: Attention Is All You Need, *CoRR*, Vol.abs/1706.03762 (2017) (online), available from ⟨http://arxiv.org/abs/1706.03762⟩.
[29] Bahdanau, D., Cho, K. and Bengio, Y.: Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[30] 国立国語研究所：現代書き言葉 UniDic (2.1.2),「UniDic」国語研短単位自動解析用辞書—最新版ダウンロード（オンライン），入手先 ⟨https://unidic.ninjal.ac.jp/download#unidic_bccwj⟩（参照 2021-04-25）.
[31] 伝 康晴，小木曽智信，小椋秀樹，山田 篤，峯松信明，内元清貴，小磯花絵：コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用，日本語科学，Vol.22, pp.101–123 (2007).
[32] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S. and Vollgraf, R.: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (*Demonstrations*), Minneapolis, Minnesota, Association for Computational Linguistics, pp.54–59 (online), DOI: 10.18653/v1/N19-4010 (2019).
[33] Alan Akbik (alanakbik): Flair, flairNLP/flair: A very simple framework for state-of-the-art Natural Language Processing (NLP) (online), available from ⟨https://github.com/flairNLP/flair⟩ (accessed 2021-04-25).
[34] Alan Akbik (alanakbik): Tutorial 4: List of All Word Embeddings, flair/TUTORIAL_4_ELMO_BERT_FLAIR_EMBEDDING.md (online), available from ⟨https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_4_ELMO_BERT_FLAIR_EMBEDDING.md⟩ (accessed 2021-04-25).
[35] Tjong Kim Sang, E.F. and Veenstra, J.: Representing Text Chunks, *9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, Association for Computational Linguistics, pp.173–179 (1999) (online), available from ⟨https://www.aclweb.org/anthology/E99-1023⟩.
[36] Stefan Schweter (stefan–it) and Alan Akbik (alanakbik): Flair Embeddings, flair/FLAIR_EMBEDDINGS.md (online), available from ⟨https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md⟩ (accessed 2021-04-25).
[37] Nakayama, H.: seqeval: A Python framework for sequence labeling evaluation, chakki-works/seqeval: A Python framework for sequence labeling evaluation (named-entity recognition, pos tagging, etc...) (online), available from ⟨https://github.com/chakki-works/seqeval⟩ (accessed 2021-04-21).
[38] Yoshiga, N.: Named Entity Recognition (NER) for Ogihan Nikki Mokuroku titles, flair-nikki-train.ipynb (online), available from ⟨https://colab.research.google.com/drive/1vrxCIx2o-4GiP8zewZO18UngEAQ6Lfnv?usp=sharing⟩ (accessed 2021-04-25).
[39] Yoshiga, N.: Named Entity Recognition (NER) for Ogihan Nikki Mokuroku titles, flair-nikki-tagger.ipynb (online), available from ⟨https://colab.research.google.com/drive/1_30rwEPSp6P5EOLzFn_glxN-wEX-Dw4I?usp=sharing⟩ (accessed 2021-04-25).

吉賀 夏子 （正会員）

1993 年佐賀大学大学院農学研究科修士課程修了，2015 年同大学院工学系研究科博士前期課程修了．2018 年同大学院工学系研究科博士後期課程修了．博士（学術）．2020 年から佐賀大学地域学歴史文化研究センター講師（研究機関研究員）．人文情報学および市民科学による地域課題の解決に関心を寄せる．デジタルアーカイブ学会，人工知能学会各会員．

堀 良彰 （正会員）

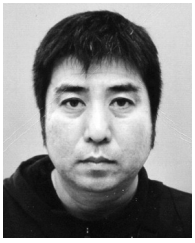
1994 年九州工業大学大学院情報工学研究科情報システム専攻修士課程修了．同年九州芸術工科大学芸術工学部助手．2004 年九州大学大学院システム情報科学研究院助教授．2013 年佐賀大学全学教育機構教授．2021 年同総合情報基盤センター教授，同センター長．ネットワークセキュリティ，ネットワークアーキテクチャ，情報システムに関する研究に従事．博士（情報工学）．ACM，電子情報通信学会，IEEE 各会員．

只木 進一 （正会員）

1987 年東北大学大学院理学研究科博士後期課程修了．理学博士．1990 年佐賀大学理工学部助教授．2000 年佐賀大学学術情報処理センター教授．2006 年同センター長．2013 年佐賀大学工学系研究科教授．計算科学，学術情報システムの管理運営技術等を専門とする．日本物理学会，アメリカ物理学会，応用数理学会各会員．

永崎 研宣 （正会員）

一般財団法人人文情報学研究所主席研究員．2000 年筑波大学大学院博士課程哲学・思想研究科満期退学．博士（文化交渉学）．東京外国語大学アジア・アフリカ言語文化研究所，山口県立大学を経て，一般財団法人人文情報学研究所の設立に参画し，現在に至る．人文情報学，仏教学の研究に従事．



伊藤 昭弘

2004 年九州大学大学院文学研究科博士課程修了．博士（文学）．2006 年佐賀大学地域学歴史文化研究センター講師．2007 年同准教授．2020 年同教授，センター長．日本近世史を専門とする．日本史研究会，社会経済史学会各会員．