Vol.2019-MPS-126 No.6
2019/12/11

近代書籍における低出現頻度文字種の獲得

藤田未希[1]　竹本有紀[1]　石川由羽[2]　髙田雅美[1]　城和貴[1]

概要：本稿では，近代書籍における低出現頻度文字種を獲得する手法を提案する．国立国会図書館で公開されている近代書籍を対象にした OCR は学習データが少ないため，認識率は十分でない．そこで，本稿では文字種の分野・領域をドメインと定義し，近代書籍における低出現頻度文字種が頻出する特定のドメインから，近代書籍用 OCR の認識率向上に必要な低出現頻度文字種を獲得する手法を提案する．まず初めに，学習データの収集対象である青空文庫の書籍の文字の出現頻度を調べ，低出現頻度文字種獲得の難易度の調査を行う．そして，分野の違う書籍として新潟県連合産婆会会報を選択し，近代書籍における低出現頻度文字種が頻出しているかを確認する．次に，青空文庫との文字の出現頻度を比較する実験を行い，提案した手法の有用性を確認する．

キーワード：近代書籍，低出現頻度文字種，文字認識，デジタルアーカイブ

Acquiring Low Appearance Characters in Early-Modern Japanese Printed Books

MIKI FUJITA[†1]　YUKI TAKEMOTO[†1]
YU ISHIKAWA[†2]　MASAMI TAKATA[†1]
KAZUKI JOE[†1]

1. はじめに

国立国会図書館[1]は，図書や雑誌，マイクロ資料等を含めるとおよそ 3,900 万点の蔵書を誇っている．これらの蔵書のうち明治期から昭和初期にかけて刊行された近代書籍は，哲学から産業，文学，芸術等幅広い分野にわたっており，現在は絶版になっているものが多い．そのため，近代の検証において学術的に貴重な資料である．そこで，国立国会図書館では，近代に刊行された書籍の中から著作権保護期間が終了したもの，または著作権者の許諾を得たものから順にデジタル化を行い，近代書籍を国立国会図書館デジタルコレクション[2]で公開している．

国立国会図書館デジタルコレクションで公開されている近代書籍は，画像データとしてアーカイブ化されているため，テキストデータは存在しない．そのため，このアーカイブの文書に含まれている文字列を検索することが出来ない．

現在のように規格化されたフォントの文書であれば，OCR によって画像からテキストデータに自動的に変換を行うことは可能である．しかし，近代書籍は活版活字印刷であるため，そこで使われているフォントは出版年代・出版者ごとに異なり，出版者数は国立国会図書館で確認できるだけで約 2 万となっている．そのため，既存の OCR では近代書籍のテキスト化は困難である．そこで，我々は近代書籍に特化した文字認識の研究に着手している [3]．近代書籍は出版者によりフォントが異なるため，複数の出版者の文字データが必要である．現在，文字データは 1 文字につき 6 者分を 2,678 種ずつ収集しており，認識率は約 97 パーセントである[4]．認識率をさらに向上させるには，1 つの文字種につき少なくとも数十者分の文字データを収集する必要があると考えられる．また，獲得する文字種の充実も不可欠である．なぜなら，既存の OCR が対象としている文字種の大半は，JIS[5]第一水準の文字種である．JIS 第二水準の文字は既存の OCR の認識率の向上には影響が小さい．しかし，近代書籍は旧字体が多く用いられていることから，JIS 第二水準の文字の出現頻度が低いとは限らない．そのため，JIS 第一・第二水準を合わせた 6,355 種のうち出現頻度上位 3,000 種程度の文字種を獲得することで，近代書籍用 OCR の認識率の大幅な向上が期待できる．しかし，現在文字収集を主に行っている書籍は，青空文庫にも登録されている近代書籍であり，1 文字につき 6 者分の文字データを揃えるのが限界である．また，収集対象の書籍では出現しない文字種もありえる．そこで，本稿では，十分な文字データを収集できていない残りの 1,000 種程度を低出現頻度文字種と定義し，近代書籍文字認識の認識率向上に必要な低出現頻度文字種を獲得する手法を提案する．

以下，本稿の構成を示す．2 章で既存研究である CNN を用いた近代書籍の文字認識について述べる．3 章では，近代書籍から低出現頻度文字種を獲得する手法の提案を行い，4 章では，文字の出現頻度調査の実験方法と結果を述べる．5 章でまとめについて述べる．

1 奈良女子大学
Nara Women's University
2 滋賀大学
Shiga University

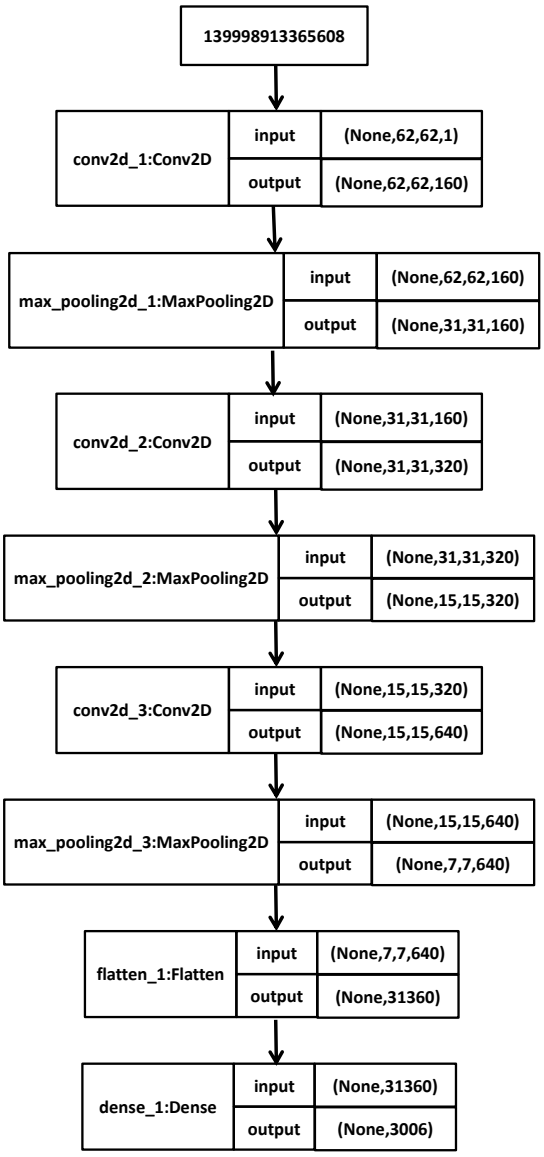

図 1　学習モデルの内部構造

2. 既存研究：CNN を用いた近代書籍の文字認識

近代書籍の文字認識には畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）[4]を使用する．CNN は，主に画像認識に利用されるニューラルネットワークの 1 つである．画像の局所的な特徴抽出を担う畳み込み層と，局所ごとに特徴をまとめあげるプーリング層を繰り返した構造となっている．図 1 は，本稿で使用する CNN のモデルである．入力には 62×62 サイズの pgm 画像を用いる．このモデルは，一般的な畳み込み計算である Conv2D を使用した畳み込み層とマックスプーリングを使用したプーリング層を 3 回繰り返した 6 層と，データを 3 次元から 1 次元に変形する Flatten 層，全結合層の 8 層で構成されている．出力部分では，ソフトマックス関数で 0 から 1 に正規化し出力する．学習データは，現代フォントと近代書籍フォントの 2 種類がある．現代フォントはヒラギノやメイリオなど 21 種類のフォントを学習している．近代書籍フォントは，青空文庫と国立国会図書館デジタルコレクションの両方で公開されている近代書籍の画像データから切り出された文字画像データであり，1 つの文字種につき 6 出版者のデータがある．文字種は 2,678 種用意されている．現在の学習データは，ひらがなと JIS 第一水準の文字種で構成されている．

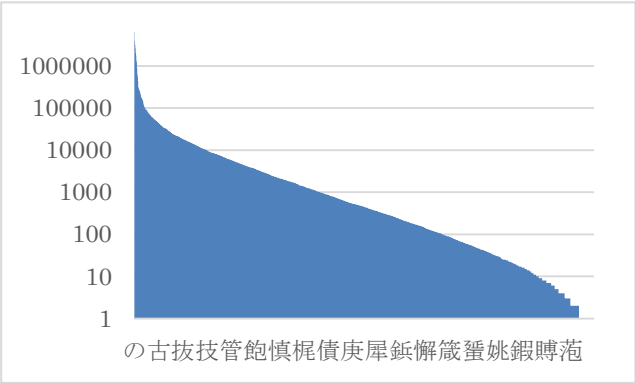

図 2　青空文庫の文字の出現頻度

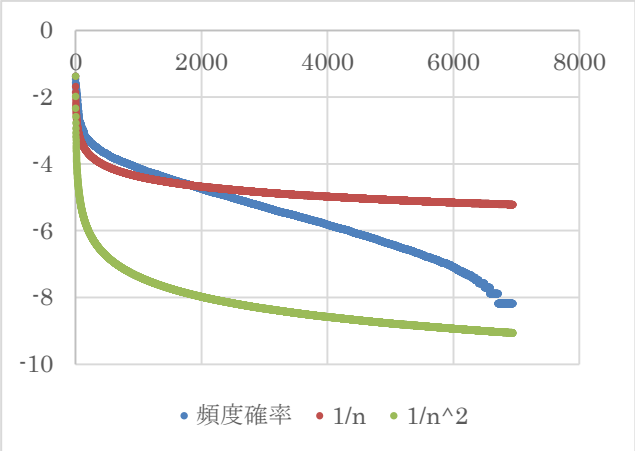

図 3　青空文庫の文字の頻度確率と $1/n$，$1/n^2$ との比較

本稿では，文字の出現頻度の調査を行う際にこの学習モデルを使用する．このモデルの現在の認識率は 97%であり，文字の出現頻度の調査で使用するには十分な認識率である．

3. 提案手法

3.1 ドメインの定義

青空文庫で公開されている書籍の文字の出現頻度分布を図 2 で示す．縦軸は文字種ごとの出現頻度の対数値，横軸は文字種である．図 2 から，青空文庫で公開されている書籍の文字の頻度分布は，高出現頻度文字種と低出現頻度文字種の出現頻度の差が大きい分布となり，文字の出現頻度に差が生じていることが確認できる．また，図 3 は青空

文庫で公開されている書籍の文字の頻度確率の散布図である. 縦軸は頻度確率の対数値, 横軸は出現頻度順位である. この文字の頻度分布は，出現頻度上位 2,000 種においてジップの法則[6]を満たしている.

ジップの法則とは，文章中で使われる単語の出現頻度を集計して頻度順に並べた際に，出現頻度がｎ番目に高い文字種が，全体に占める割合はｎ分の 1 に比例する法則である. 例えば, 英語の書籍の英単語の出現頻度を調べた場合, 出現頻度上位が以下の様になったとする.

（1） the（全体の 10%）
（2） of（全体の 5%）
（3） and（全体の 3.3%）

「the」や「of」など極一部の単語の登場回数が圧倒的に多く, 残りの大半の言葉は 1, 2 回の低頻度しか登場しない. このとき, 単語の出現率と順位の積はおよそ 10 の定数になる．このことから，出現頻度の順位が低い文字種ほど出現回数は極めて少ないことが確認できる．また，図 3 から青空文庫で公開されている書籍の文字は, 出現頻度上位 2,000 位以下の文字種においてｎ分の 1 よりも低く，n^2 分の 1 よりも高い獲得確率であると考えられる．すなわち，出現頻度 2,000 位以下の低出現頻度文字種はジップの法則すら満たしておらず，発見することが極めて困難であることが分かる.

青空文庫で公開されている書籍では，「彼」や「曰」など口語的な表現で使用される文字種の出現頻度が圧倒的に多く，「砿」や「碕」など専門用語で使用されることが多い文字種の出現回数は 1 回程度である．以上から，文字の出現頻度の差を満たすためには，文字種の出現頻度分布が異なる書籍から獲得する必要がある．本稿では，文字種の分野・領域をドメインと定義する.

ドメインを変更して文字の出現頻度を調べた場合，医学の分野においては「療養」や「医療」, 数学の分野においては「幾何」や「微分」, 文学の分野においては「今日」や「山」など，ドメインごとに出現頻度上位の文字種は異なると考えられる．また，同じ分野の内容を述べる際に多くの場合同じ文字が使用される．従って，文字の頻度分布が似ている書籍は，分野・領域が類似した内容であることが期待できる．実際に，昭和初期に刊行された愛媛県農業史[7]の資料には，藩政時代の農業の状況について記載されているため,「藩」や「班田制」などの歴史分野に関する単語が確認できた．また, 農業に関する用語である「田畝」や「耕地」など文学の分野に属する近代書籍ではあまり見かけることのない単語が頻出している．以上から，ドメインごとに文字の出現頻度分布は異なり，出現頻度の高い文字種も違うことが示唆される．従って，国会図書館デジタルアーカイブの近代書籍における低出現頻度文字種が頻出する近代書籍も存在する可能性がある．多くのドメインから高出現頻度文字種を獲得することで，獲得できる文字種の充実が期待できる.

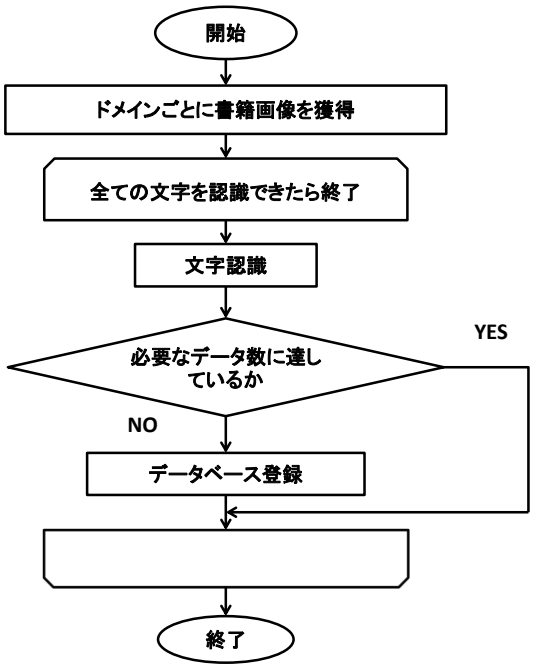

図 4　提案手法フローチャート

3.2 低出現頻度文字種獲得のための提案手法

本稿では，ドメインごとに文字収集を行い，近代書籍における低出現頻度文字種が頻出する文字種の集合から必要な文字種を獲得する手法を提案する．提案手法では，本研究室で開発を行っている近代書籍 OCR 支援 Web アプリケーション[8]内のデータベースを使用する．近代書籍 OCR 支援 Web アプリケーションは，近代書籍用 OCR の認識率向上に必要な学習データを効率良く収集することを目的に開発が行われている. 先に述べたように, 近代書籍は時代・出版者によりフォントが異なるため，実用レベルの認識率を得るためには，1 つの文字種につき数十者分の文字データが必要である．しかし，現在文字収集の対象である青空文庫と国立国会図書館デジタルコレクションの両方で公開されている近代書籍のドメインは文学であるため，すべての低出現頻度文字種につき数十者分の文字データを獲得することは不可能である．そこで，本稿では近代書籍から低出現頻度文字種を獲得する手法を提案する．提案手法のフローチャートを図 4 に示す．まず初めに，ドメインごとに認識で使用する書籍画像を獲得する．次に，獲得した書籍画像で文字認識を行う．この後，近代書籍用 OCR の認識率向上に必要な低出現頻度文字種の文字収集を行う．データベースに登録する際に，認識した文字が必要なデータ数に達しているかの確認をする．必要なデータ数に達してい

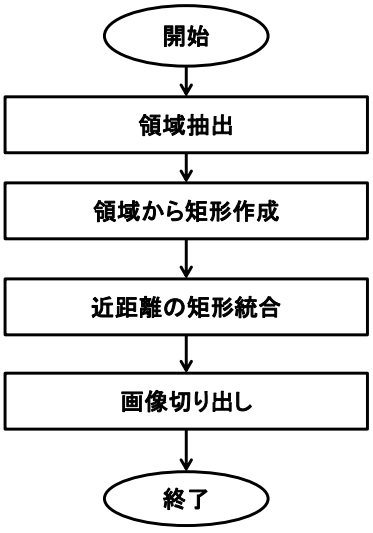

図 5　文字切り出し処理のフローチャート

るかの判定は，データベースに登録されている文字種の検索で判定を行う．判定基準は，認識した文字の文字種がデータベースに登録されていない文字種である，または認識した文字の文字種はデータベースに登録されているが年代･出版者が異なるフォントであることを判断基準とする．必要なデータ数に達していない文字種である場合，データベースに登録する処理を行う．認識したすべての文字をデータベースに登録する方法の場合，必要なデータ数に達している文字種もデータベースに登録する処理を行うため，時間がかかり非効率である．提案手法では必要な文字種のみ登録処理を行うので，膨大な量の書籍から低出現頻度文字種を効率良く獲得することが可能である．

4. 実験

4.1 実験方法

本章では，近代書籍と特定のドメインの文字の出現頻度の比較実験を行う． 実験手順について説明する．まず，近代書籍として青空文庫で公開されている書籍の文字の出現頻度を調査する．次に，特定のドメインに出現する文字の出現頻度を調査する．特定のドメインの文字認識には既存研究[5]で紹介した CNN を用いる．文字認識の手順は以下の通りである．

手順1. 文字切り出し
手順2. 切り出した文字画像の正規化
手順3. CNN で認識
手順4. 認識結果の確認・修正

手順 1 では認識で使用する書籍画像の文字切り出しを行う．図 5 は，認識の前処理である文字切り出し処理のフローチャートである．まず初めに，領域抽出処理で細かいパーツの領域を検出する．検出する領域は，画像内の要素のうち外側にある輪郭である．次に，検出した領域を囲む矩形を作成し，矩形が重なる場合や距離が近い場合，矩形を統合する．そして，作成した矩形で画像を切り出して文字切り出し処理を行う．次に手順 2 で，切り出した文字画像の正規化を行う．手順 2 では主にノイズ画像の削除と画像サイズと拡張子の変更を行う．ノイズ画像とは，書籍画像から文字を切り出した際に文字部分ではなく，書籍画像の汚れなどを文字と認識して切り出された画像である．ノイズ画像は文字画像と比べデータサイズが小さいので，一定のデータサイズより小さい画像を一括削除するプログラムを用いる．同時に，画像サイズは 62×62，拡張子は pgm 形式に変更する．CNN を用いた文字認識を手順 3 で行う．学習モデルは，既存研究[4]で紹介したものと同じものを使用する．最後に手順 4 で認識結果の確認を行う．認識結果は，認識した文字画像・認識結果（文字）・スコアを表示する．スコアは認識結果の信頼度を意味する．スコアが低いものを目視で確認し，誤認識していた場合，手作業で修正を行う．近代書籍と特定のドメインの文字の出現頻度をそれぞれ調べた後，特定のドメインにおいて出現頻度上位の文字種が近代書籍では出現頻度上位何位になるのか確認する．出現頻度上位文字種の確認方法は，Microsoft Excel の MATCH 関数を使用する．青空文庫で公開されている書籍の文字の出現頻度を出現回数の多い順に登録した Excel シートにおいて，特定のドメインの出現頻度上位の文字種をそれぞれ MATCH 関数で検索する．本実験では，特定のドメインにおいて出現頻度上位 15 位の文字種の調査を行う．

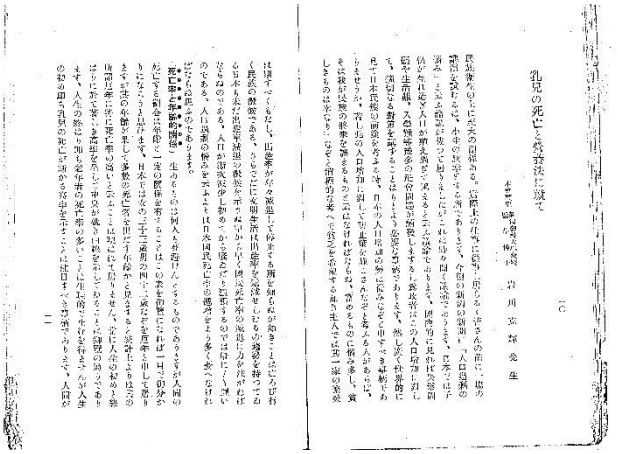
乳兒の死亡と堕胎法に就て

図 6　新潟県連合産婆会会報創刊号

4.2 実験データ

本実験で使用する実験データについて述べる．本実験では，近代書籍として文学作品を扱っている青空文庫を用いる．青空文庫は，git hub で公開されているテキストデータを使用する[9]．作品数は 10,428 タイトルである．特定のド

表 1　文字種の割合（近代書籍）

	文字種	割合
総文字種	6,646 種	100%
JIS 第一水準	2,964 種	45%
JIS 第二水準	3,121 種	47%
その他	561 種	8%

表 2　文字種の出現頻度の割合（近代書籍）

	総出現回数	割合
総文字種	153,668,874	100%
JIS 第一水準	41,867,457	27%
JIS 第二水準	952,891	1%
その他	110,848,526	72%

表 3　出現頻度上位文字種の比較結果

文字種	近代書籍（位）
會	1520
一	46
產（産）	なし（63）
合	142
婆	977
乳	1369
議	507
長	178
組	657
事	80
兒	2476
出	57
生	88
其	263
總	2716

メインは，日本産婆学会[10]とする．特定のドメインの実験データは，新潟県連合産婆会で刊行された新潟県連合産婆会会報創刊号を使用する．

近代書籍の文字の出現頻度のデータは，2 章で調査したものを使用する. 資料は Microsoft Excel で作成されている．新潟県連合産婆会会報の資料は，奈良女子大学の松岡教授（医療人類学）より提供されたものである．資料は紙で提供されているため，1200dpi の白黒画像でスキャンを行いデジタルデータに変換を行った．図 6 は，昭和 3 年に刊行された新潟県連合産婆会会報創刊号の資料である．新潟県連合産婆会会報創刊号では，乳児の死亡率と栄養法について，死亡率と年齢的関係や人乳と牛乳の栄養分の観点からまとめられている．この資料は講演会の議事録であり，本文は口語的文章で記載されている．産婆会は，地方独自の情報共有を目的に戦前に設立された[10]．このことから，書式やフォントも地方によって大幅に違うことが予想される．本稿では，新潟県の日本産婆学会の資料を使用する．出版者は清水印刷所である．

4.3 実験結果と考察

本実験では，近代書籍の総文字出現回数は 153,668,874 であり，6,646 種の文字種が確認できた．表 1 と表 2 は近代書籍の文字種の割合と出現頻度の割合である．その他には記号・ひらがな・カタカナが含まれている．JIS 第一水準･JIS 第二水準･その他の判定は, 変換した JIS コードが, 12321 以上かつ 20307 以下である場合 JIS 第一水準，20513 以上かつ 29734 以下である場合 JIS 第二水準，それ以外の場合をその他とする．表 1 と表 2 から，JIS 第一水準と JIS 第二水準の出現する割合は同じであるが，JIS 第二水準の出現頻度は極めて低く，ひらがな・カタカナの割合が大半を占めていることが分かる．

表 3 は, 新潟県連合産婆会会報創刊号の出現頻度上位 15 位の文字種を近代書籍における出現頻度と比較した表である．ひらがな文字は両方のドメインで頻出するため，表 3 では漢字のみの結果を載せている．本文は議事録も記載されているため,「事」や「議」,「長」,「其」, の議事録で使用されることが多い文字種は，日本産婆学会と近代書籍の両方で出現頻度が高い．さらに，新潟県連合産婆会会報創刊号では「産」や「婆」,「乳」など妊娠や産婆に関する文字種が頻出しており，これらの文字種は，近代書籍では出現頻度があまり高くないことが表 3 から読み取れる．実験データが，乳児の死亡率と栄養法など産後の妊婦を対象とした内容を記載している資料であったため，このような結果になったと推測される．

昭和 3 年に刊行された新潟県連合産婆会会報創刊号は,「産」の旧字体である「產」や「児」の旧字体である「兒」,「会」の旧字体である「會」,「総」の旧字体である「總」など旧字体の出現頻度が高いことも確認できる. 表 3 から,「産」は近代書籍では使用されておらず，他の旧字体の出現頻度も低いことが分かる．また，新潟県連合産婆会会報創刊号において出現頻度上位 15 位の文字種ではないが, 出現頻度の高かった「県」の旧字体である「縣」の近代書籍での出現頻度は 3118 位であり,「妊」の異字体である「姙」の近代書籍での出現頻度は 3842 位であった. この結果から, 刊行されている近代書籍全般では旧字体が頻出する可能性があることが示唆される．

今後は，あらゆるドメインから近代書籍における低出現頻度文字種を獲得し，近代書籍用 OCR の認識率向上を目指す．

5. まとめ

本稿では，近代書籍用 OCR の認識率向上に必要な低出現頻度文字種を獲得する手法を提案し，文字収集を容易に行

うことを目指している．まず初めに，青空文庫で公開されている書籍の文字の出現頻度分布を調査した．その結果，高出現頻度文字種と低出現頻度文字種の出現頻度には大きな差があることが確認できた．そして，青空文庫で公開されている書籍の文字の頻度確率は，出現頻度上位 2,000 位の文字種においてジップの法則を満たしているが，出現頻度上位 2,000 位以下の低出現頻度文字種は，ジップの法則のｎ分の 1 よりも低く，n^2分の 1 より高い獲得確率であることが分かった．このことから，低出現頻度文字種の発見は極めて困難であり，異なる頻度分布を持つ分野の書籍からの文字収集が必要であると考えた．書籍の分野・領域を変更して文字の出現頻度を調べた場合，専門用語など文学の分野・領域ではあまり見かけない文字種が多く出現することから，文字の出現頻度が異なることが推測される．従って，近代書籍における低出現頻度文字種が頻出する書籍が存在することが示唆される．

本稿では文字の分野・領域をドメインと定義し，近代書籍から低出現頻度文字種を獲得する手法を提案する．提案する手法は，ドメインごとに書籍画像を獲得し，文字収集を行い，必要なデータ数に達していない文字のみデータベースに登録するという収集方法である. 提案する手法では，膨大な量の書籍から近代書籍における低出現頻度文字種を効率良く獲得することが可能になる．

実験では，近代書籍（青空文庫）と特定のドメイン（日本産婆学会）の文字の出現頻度の違いを確認した．その結果,「総会」の旧字体である「總」や「會」など議事録で使用される文字種や,「乳児」や「産婆」など出産に関する用語の文字種が，新潟県連合産婆会会報創刊号では出現頻度が高く，近代書籍では低出現頻度文字種であることが分かった．また,「産」の旧字体である「產」は近代書籍では出現せず,「兒」や「總」など他の旧字体も，近代書籍では出現頻度が高くないことが確認できた．実験結果から，ドメインごとに出現頻度の高い文字種は異なり，近代書籍における低出現頻度文字種が頻出するも書籍が存在することが確認できる. また, 文字の出現頻度を比較することにより，近代書籍における低出現頻度文字種を獲得できることが判明した．今後は，あらゆるドメインから近代書籍における低出現頻度文字種を獲得し，近代書籍用 OCR の認識率向上を目指す．

謝辞　本研究は文部科学省科学研究費補助金（17H01829）の助成を受けたものである．
また，新潟県産婆会に関する資料は奈良女子大学の松岡悦子教授より提供されたものである．

参考文献

[1] 国立国会図書館：http://www.ndl.go.jp/（参照:2019/11/05）
[2] 国立国会図書館デジタルコレクション：http://kindai.ndl.go.jp/（参照:2019/11/05）
[3] Fujimoto,K.，Ishikawa,Y.，Takata,M. and Joe,K.：Early-Modern Printed Character Recognition using Ensemble Learning，Processing of The 2017 International Conference on Parallel and Distributed Processing Technologies and Applications, Vol. Ⅰ, pp. 288-294(2017).
[4] Yasunami,S.，Takemoto,Y.，Ishikawa,Y.，Takata,M. and Joe, K.：Applying CNNs to Early-Modern Printed Japanese Character Recognition，Proceedings of The 2019 International Conference on Parallel and Distributed Processing Technologies and Applications.
[5] 日本工業標準調査会：http://www.jisc.go.jp/index.html（参照:2019/11/05）
[6] Zipf,G. K.(1949): Human Behavior & The Principle of Least Effort, An Introduction to Human Ecology, Addison-Wesley Press Inc..
[7] 愛媛県農業史．上巻: http://dl.ndl.go.jp/info:ndljp/pid/1066451（参照:2019/11/05）
[8] Kosaka, K., Awazu, T., Ishikawa, Y., Takata, M, and Joe, K.: An Effective and Interactive Training Data Collection Method for Early-Modern Japanese Printed Character Recognition, Proceeding of The 2015 International Conference on Parallel and Distributed Processing Techniques and Applications, Vol.1, pp. 276-282 (2015)
[9] 青空文庫：https://github.com/aozorabunko/aozorabunko（参照:2019/11/05）
[10] 日本助産師会：http://www.midwife.or.jp/index.html（参照:2019/11/05）