くずし字の翻刻に向けたディープラーニングの活用と分析*

小島 朋佳** 植木 一也**

Utilization and Analysis of Deep Learning for *Kuzushiji* Translation

Tomoka KOJIMA and Kazuya UEKI

In this study, to promote the translation and digitization of historical documents, we attempted to recognize Japanese classical *kuzushiji* characters by using the dataset released by the Center for Open Data in the Humanities (CODH). Using deep learning, which has undergone remarkable development in the field of image classification, we analyzed how successfully deep learning could classify more than 1,500-class *kuzushiji* characters through experiments, and what made it difficult to classify *kuzushiji* characters. In addition, we introduced a method to automatically eliminate characters that were difficult to classify or characters that were not used during training. In an actual translation work, those ambiguous characters can be left as unknown characters as “= (geta)” and pass on to the expert to make a decision. Finally, our experiments showed that the classification rate was improved from 72.10% to over 90% by performing the data augmentation and classifying only characters with high confidence and confirmed the effectiveness of our proposed method.

Key words: Japanese classical *kuzushiji* characters, character recognition, deep learning, data augmentation, unknown class, OCR

1. はじめに

日本の文字は時代の流れとともに変化しており、今日、日本の古典文学を専門家以外が読むことは困難である。歴史的文献を解読することで、その時代に何が行われてきたかを知ることができるため、歴史的文献の翻刻の研究や電子化が進められている。しかしながら、いまだ多くの文学作品が電子化されずに埋もれてしまっているという現状がある。この電子化の障壁となっているもののひとつが、古典文学が現代人に容易に理解できないくずし字¹で書かれていることである。くずし字の認識を難しくしている要因は以下のとおりである。

- 古典籍で使われるくずし字は手書き文字であるため、筆者によって文字の書き方がさまざまであること
- つづけ字²が多いことから文字画像の切り取りが困難であること
- くずし字はバリエーションが多いため、同じ文字でも見た目が大きく異なり、他の文字とのはっきりとした区別が困難であること

この問題に対し、ディープラーニングによるくずし字認識により、文学作品を電子化して翻刻の利便性を向上させる取り組みがされている。現代の手書き文字に対する認識はディープラーニングの技術の発展により、高い精度で認識することが可能となっている。

本稿では、現代の文字で高い認識精度を示すディープラーニングをくずし字の認識に適用し、どの程度の精度が得られるかを検証する。さらに認識結果から詳細の分析をして、くずし字認識の課題とそれに向けた問題解決方法を検討することを目的とする。また、実際の翻刻の場面についても想定し、認識が困

難な文字の取り扱いについても検討して、翻刻に役立てる。

本研究の貢献は以下の3点である。

- 従来のくずし字認識の研究では、主に50種類程度のひらがなのみの実験が行われていたのに対し、漢字やカタカナなどを含めた1,500種を超える文字を認識を試み、どのような問題点があるかを示したこと。
- 実際の翻刻作業では、今までに翻刻されていない作品を取り扱うことから、学習で使用するデータとテストで使用するデータでは、全く異なる作品・作者のものをを用いるという難しい環境下で実験し、分析を行った点。
- 識別困難な文字を保留して後工程の専門家に委ねる基準として、ニューラルネットワークの最大の出力値を採用し、その有効性を示した点。

以下、2章では先行研究について述べ、3章では実験方法について説明する。4章ではディープラーニングを用いたくずし字の認識実験を示し、5章で本研究のまとめと今後の課題を述べる。

2. 先行研究

近年、書籍の電子化に伴いOCR技術も進展し、現代の活字文書の解読正解率は99%を超える精度を達成している。先行研究においても、欠損のない文字画像に対しては比較的良好な認識精度が得られている。しかし、古文書にはノイズや欠損のある文字画像があり、そのことが認識率の低下を招くおそれがある。

機械学習の技術を用いてくずし字を認識する試みがいくつか報告されている¹⁾²⁾。文献¹⁾では、ディープラーニングを用い

* 原稿受付 平成31年4月23日

掲載決定 令和元年8月21日

** 明星大学 情報学部 情報学科 (東京都日野市2-1-1)

¹ 本論文では、くずし字は江戸時代以前の古文書に書かれている草書の文字のことを指す。

² つづけ字とは、一字一字を区切らずに続けて書く文字のことをいう。

て、48 種類のひらがなを認識する学習モデルを作成し、WWW アプリケーションとして実装している。文献²⁾においても、ディープラーニングを用い、ひらがなのくずし字 46 種類を認識する実験結果が報告されている。ディープラーニングの手法に加え、文字によってアスペクト比が異なることを利用し、ひらがなのくずし字 46 種類に対して、約 90% の精度を達成した。文献³⁾では、2017 年の PRMU アルゴリズムコンテスト^{*3}のくずし字認識チャレンジの成果が報告されている。コンテストでは、漢字を除く変体かなを認識対象として、難易度によって 3 つのレベル分けをしている。レベル 1 では 1 文字ずつ、レベル 2 は縦方向の 3 文字、レベル 3 は縦横方向の 3 文字以上の文字が認識の対象である。上記のように、くずし字認識の研究は、ひらがなの認識を中心として行われてきているが、古典籍には漢字やカタカナなどひらがな以外の文字も含まれている。古典籍を電子的に読めるようにするためには、広範囲の文字を認識しそれぞれの認識精度を高めていく必要がある。

文献⁴⁾では、ディープラーニングを用いて、1895 年から 1945 年に記録された台湾総督府の文書の文字の認識を試みている。この研究では、3000 を超える漢字を含む文字の種類を扱い、クラス間のデータ数がアンバランスな問題を、データ拡張により解決している。

手書き文字の認識として、Omniglot データセットを用いた実験も報告されている⁵⁾。このデータセットは 1623 種の各文字につき 20 枚と比較的データ数の少ないデータのため、ワンショット学習の実験に活用されている。この論文では、人間は物事を要素に分解して考えるという特性と同様に、文字を書く過程を分解して書き方の常識を学ぶことを提案している。実験では、Bayesian program learning を用いたワンショットでの文字の識別を行っている。これはクラス内のデータ数が少ない場合に有効である。しかしながら、本稿で扱うくずし字のデータセットは、クラスの中で文字の書き方のバリエーションが多いことから、データ数が多いクラスについては、ディープラーニングの手法の方が優れていると考えられる。

また、文献⁶⁾では、負荷の高い翻刻の省力化を目的に、新方式 OCR 技術を提案している。この技術では、完全自動処理ではなく、専門家と非専門家および自動処理の分業により、翻刻作業の省力化を目指している。自動処理の OCR のみで、100% の解読精度は目指さず、類似画像検索システムにより類似度が基準に満たない文字を「＝ (ゲタ)」として残し、後工程の人 (専門家) に判断を委ねることで、結果的に速く、高精度な翻刻ができると述べられている。

以上のことから本研究では、ディープラーニングを用いて漢字を含む多くの文字種の識別を試みることに加え、ニューラルネットワークの最大の出力値が低い文字をあえて識別せずに「＝ (ゲタ)」とすることを提案する。特に、誤って識別している文字や、学習時になかった文字を自動的に排除できるかについて実験により検証した。

3. 実験内容

3.1 日本古典籍字形データセット

本研究では、人文学オープンデータ共同利用センターが公開している日本古典籍字形データセット⁷⁾を利用する。日本古典籍字形データセットには「おらが春」、「雨月物語」などの有

Fig.1 The number of images in each class in the kuzushiji dataset⁷⁾

Fig.2 Examples of similar shapes despite characters being different. Left: Character “り” (Unicode: U+308A) was originally derived from jibo “利”. Right: Character “わ” (Unicode: U+308F) was originally derived from jibo “利”

Fig.3 Examples of separate deformation caused by different jibo. Various shapes for the one character “あ” (Unicode: U+3042). Jibo on the left is “安”, and jibo on the right is “阿”

Fig.4 Examples of differences in deformation despite being linked to the same jibo depending on the work or the writer. The character “な” (Unicode: U+306A) was originally derived from the same jibo “奈”

名な物語や「料理物語」など江戸時代の食文化がわかる作品等、江戸時代の 28 作品が含まれている。本研究では、これらの作品からひらがな、カタカナ、漢字の文字の領域を切り取られた画像を用いた。2019 年 3 月時点では、4,645 文字種の字形データ 684,165 文字のくずし字画像を利用することができる。図 1 に文字のクラスごとの画像データ数を示す。最も画像数が多いクラスは「に」(文字コード: U+306B)で、24,710 枚あるのに対し、多くのクラスは画像が極端に少なく、画像が 1 枚しかないクラスも 882 種類存在するなど、クラスごとにデータ数の偏りが大きい。

データベース中の画像を確認すると、文字の種類が多様であ

^{*3} 研究会活動の活性化を目的としてパターン認識・メディア理解研究会 (PRMU) が毎年主催しているコンテスト。

Fig.5 Configuration of our neural network model

ることから、図2に示すように異なる字種にも関わらず似た字形をもつ文字があり、はっきりとした区別が困難なものが多く存在していた。また、図3に示すように、同一文字にも関わらず見た目が大きく異なる文字も存在していた。これは、ひらがなとカタカナのくずし字は、字母と呼ばれるくずし字のもととなっている漢字があり、複数の字母がもとになっている同一文字が存在することに起因する。さらに、図4に示すように同一文字、同一字母であっても作品や筆者によって文字のくずれ方が異なる文字も存在していた。

3.2 実験で使用するデータの分割

本研究では、日本古典籍字形データセットの28作品のデータを学習データ、検証データ、テストデータに分けて実験する。学習データは認識モデルの学習に使用し、1エポックごとに検証データを用いて評価し、最適なモデルを選択する。最終的に選択されたモデルでテストデータを評価し、モデルの性能を測る。具体的には、「おらが春」と「雨月物語」をテストデータとし、残りの26作品を学習データと検証データとした。26作品の画像データ数は628,136枚でクラス数は4,493となった。

クラスごとのデータ数の偏りが大きいため、本研究の実験では、26作品の画像データのうち、20枚以上の画像が利用できる1,506の文字種に限定した上で実験した。画像データを限定したところ、利用できる画像データ数は全部で613,313枚となった。学習データと検証データは8:2の比率で分けたため、学習データは490,650枚、検証データは122,663枚となった。

3.3 使用するニューラルネットワークの構造

今回使用するデータベースを提供している人文学オープンデータ共同利用センターでは、ディープラーニングを用いたPythonの文字認識のサンプルプログラムも提供しており、容易にくずし字の認識実験をすることができる。本実験でも、提供されているプログラムとディープラーニングの構造をそのまま用いた。ここで利用する畳み込みニューラルネットワーク（以下、CNN）のモデルは、MNISTの文字認識のために作成されたもので、図5に示すような畳み込み層（プーリング層を含む）2層と全結合層1層を含む構造になっている。すべての画像はグレースケール画像に変換するとともに、28×28ピクセルの大きさにリサイズしてCNNに入力した。活性化関数にはReLU (Rectified Linear Unit) を、出力にはSoftmax関数を用い、それぞれの文字に分類される確率を出力する。最適化手法にはAdaDeltaを用いた。

4. くずし字の認識実験

この章では、これらの認識困難なくずし字をディープラーニングを用いて実験し、どの程度認識可能か、また誤識別の原因は何かについて検証する。

Fig.6 Number of training images in each character class

Fig.7 Classification rate for each character class

4.1 ベースライン認識精度の確認

くずし字のベースライン認識精度を確認するため、3章で示した日本古典籍字形データセットのうち、26作品を学習に用い、「おらが春」と「雨月物語」の2作品をテストデータとして認識実験を行った。クラスごとの学習画像データ数を図6に示す。多くのクラスは学習に利用できるデータ数が極端に少ないことから、20枚以上の画像が利用できる1,506クラスに限定した。テストデータは全部で56,029枚あり、文字のクラス数は2,249である。テストデータには、学習に使用した26作品に含まれない文字や、学習の際に排除された20枚未満のクラスの文字が合わせて3,003枚あるため、実際に識別可能な画像は53,026枚となる。

学習は120エポックまで行い、検証データで識別精度を評価して最適なモデル（本実験では、89エポックまで学習したときのモデル）を選択した。このモデルでテストデータを評価したところ、正しく識別できた画像は40,395枚であり、識別率は72.10%（= 40,395 / 56,029）となった。また、識別可能な画像のみに限定すると、識別率は76.18%（= 40,395 / 53,026）となる。図7にクラスごとの識別精度を示す。この図から学習データ数が少ないクラスの多くは正しく識別できていないことが確認でき、実際に各クラスの精度の平均は27.37%と極端に低かった。

Table1 Classification rates of different data augmentation methods

Data augmentation method	image accuracy %	class average accuracy %
Without data augmentation	76.18	27.37
Rotation	76.91	41.45
Width shift	76.85	39.86
Height shift	76.99	36.62
Channel shift	77.08	40.05
Shear	77.02	38.95
Zoom	77.18	36.05

Table2 Classification rates depending on the minimum number of images in one class

The minimum # of images in one class	image accuracy %	class average accuracy %
Without data augmentation	76.18	27.37
50	76.57	27.38
100	77.23	37.22
200	79.73	56.60
500	80.98	66.00
1000	80.20	69.80

Fig.8 Example of data augmentation. The original images are located on the upper left. The other nine images are created by data augmentation **Fig.9** Classification rate for each character class after data augmentation

4.2 データ拡張による精度向上施策

学習データ数が少ないクラスは識別精度が低くなってしまうことから、一般物体認識などのタスクでよく使われるデータ拡張をすることで、識別率の向上を試みた。具体的には、以下のデータ拡張方法を用いて変更した画像を作成し、学習データの増強を図った。

- 回転（最大で3度）
- 水平方向のシフト（最大で画像サイズの3%）
- 垂直方向のシフト（最大で画像サイズの3%）
- RGBの値を変更（最大で50）
- シアー変換（最大で $\pi/4$ ）
- ズーム（最大で画像サイズの3%）

Fig.10 Example images of character “經” in the test set (Unicode: U+7D93). Four images were mis-classified as the character “經”

まず、データ拡張手法の種類別に効果を確認する実験を行った。クラス内のデータ数が100未満のクラスに対し、各データ拡張方法を用いて変更した画像を作成し、データ数が100枚となるようにした。識別可能な画像のみに限定した場合の識別率を表1に示す。いずれのデータ拡張手法も画像単位の識別率、クラス平均識別率が向上することが確認できた。そのため、以後の実験で使用する学習データベースは、すべてのデータ拡張手法を同時に複数用いることとした。データ拡張の画像の例を図8に示す。

今回使用するデータベースは、図6に示すように、学習データ数のばらつきが極端なため、クラス内の最小データ数を変化させることで、識別精度が向上するかどうかの確認をした。クラス内の最小データ数を50, 100, 200, 500, 1000と変化させ、識別可能な画像のみに限定した場合の精度により比較をしたところ、表2のようになった。画像単位での識別率はデータ数が増すにつれて上昇し、データ数500のときに最も高くなった。一方、クラス平均識別率はデータ数が増すにつれ、27.37%から69.80%と42.43ポイントと大幅に向上した。以後の分析では、各クラスの最小データ数が500のモデルを用いることにする。データ拡張後の画像の枚数は1,156,040枚となり、これを学習データ(924,832枚)と検証データ(231,208枚)に8:2に分割した。認識モデルの学習は120エポックまでを行い、最適なモデル(112エポックの時点のモデル)を選択した。

このモデルでテストデータを評価した結果、正しく識別できた文字はデータ拡張なしの40,395枚から42,942枚に増え、識別率は72.10%(=40,395/56,029)から76.64%(=42,942/56,029)に向上した。識別可能な文字だけに限定した場合の精度は80.98%(=42,942/53,026)である。図9に、クラスごとの認識精度を示す。学習データ数が少ないクラスの認識に改善が見られた。

4.3 エラーの分析

データ拡張後の認識モデルを用いても認識精度が向上しなかった文字について、エラーを分析したところ、以下のような原因が見つかった。

- 異なる字種にも関わらず似た字形が存在すること
- ひらがなは字母によって崩れ方が異なることからバリエーションが多く、他の文字をひらがなに誤識別すること
- 筆者によって字形が大きく異なること
- 踊り字^{*4}と呼ばれる文字には類似している文字が多いことや、ボケ欠けを含む画像があること

以下、その具体例を示す。

まず、異なる字種にも関わらず似た字形が存在する例を図10に示す。文字「經」(文字コード: U+7D93)の識別率は9.1%(=3/33)と低い精度であった。この文字では図11に示すよう

^{*4} 踊り字とは、同じ漢字やひらがなやカタカナを繰り返すときに用いる符号のことをいう。踊り字にはいくつかの種類が存在し、前にある文字によって後にくる字形が異なるという決まりがある。

Fig.11 Example images of character “経” in the training set (Unicode: U+7D93)

Fig.12 Example images of character “経” in the training set (Unicode: U+7D4C)

Fig.13 Example images of character “茶” in the test set (Unicode: U+8336)

Fig.14 Example images of character “茶” in the training set (Unicode: U+8336)

Fig.15 Example images of character “ゝ” in the test set (Unicode: U+309D).

Left: two characters on the left were mis-classified as character “ゝ”.
Right: two characters on the right were mis-classified as character “く”

に、さまざまなバリエーションの文字を学習させている。識別結果を確認したところ、図 12 に示す文字「経」（文字コード：U+7D4C）に誤ってしまうエラーが多くが見られた。たとえば、図 10 の 4 枚は文字「経」と誤って識別している。

ひらがなは字母によって崩れ方が異なることからバリエーションが多く、他の文字をひらがなに誤識別する例も多く見られた。たとえば、文字「左」（文字コード：U+5DE6）は文字「る」（文字コード：U+308B）に誤って識別しているものが多く見られた。そのため、文字「左」（文字コード：U+5DE6）の識別率は 8.7%（= 4 / 46）と低い精度であった。

筆者によって字形が大きく異なる例を図 13 に示す。文字「茶」（文字コード：U+8336）の識別率は 4.2%（= 2 / 48）であった。これに対応する学習データは、図 14 に示すように、「浮世風呂」、「当世料理」、「菓子話船橋」、「膳部料理抄」、「好色一代男」など、さまざまな作品の文字が含まれており、それらは筆者によって文字のくずれ方が異なる。また、学習に用いた文字画像とテストしたものの字形が異なり、くずし字は手書きであり筆者によって字形が異なるため、識別が困難と考えられる。

さらに、踊り字が原因で識別率が低くなる例を示す。図 15 に示すような踊り字と呼ばれる文字の画像をいくつか見たところ、類似している文字や画像のボケ欠けが多いことから、識別が困難であることがわかった。ボケた字や欠けた字の画像が多く見られたのは、くずし字は複数の文字が連続して繋がって書かれているため、文字を一文字ずつ切り出すことで字が欠

Fig.16 Classification accuracy and the number of images estimated versus probability threshold

けてしまったのが原因である。たとえば、図 15 の文字「ゝ」（U+309D）の識別率は 57.6%（= 354 / 615）であった。この文字は文字「ゝ」（文字コード：U+30FD）や「く」（U+304F）に多く誤識別が見られた。これはいくつか存在する踊り字同士の形が似ていることや、踊り字に似た字形のひらがながあることが原因と考えられる。複数ある踊り字など類似した文字を区別するためには、文字を一文字ずつ識別するのではなく、連続した文字を同時に認識するなど、文脈からの判断が必要となる。

4.4 認識が困難な文字の除外

テストデータとして使用した「おらが春」、「雨月物語」の中の文字には、学習データ（26 作品）に含まれない文字が存在する。また、学習データの中で画像の枚数が 20 枚未満の文字クラスは学習の際に使用していないため、テストデータの中には学習に使われない文字も含まれる。テストの際には、学習されていない未知の文字の有無はわからないため、それらを自動的に排除する仕組みが必要となる。本実験では、テストデータ 56,029 枚のうち 3,003 枚が学習に使われていない文字である。

また、データ拡張による改善後のモデルであっても、識別率 76.64% に留まり、実際に翻刻の作業を行う場合に作業効率の低下を招いてしまう。そのため、誤った認識を可能な限り排除し、正しい認識結果のものだけを残すことが重要である。

一方、前節で述べたエラーの分析により、学習に含まれない文字や、誤識別している文字をニューラルネットワークに入力した際、その出力が不安定となり、Softmax 関数で出力される最大の確率の値が低い傾向があることがわかった。以上のことから、ニューラルネットワークの最大の出力確率（以後、確信度と呼ぶ）が高いもの、つまり認識モデルが確信を持って認識したもののみを用いることで、学習に含まれない文字や、識別誤りの文字を排除可能かを確認した。具体的には、確信度をしきい値とし、しきい値未満の文字を不明な文字として保留しておく。これは文献⁶⁾に述べられているように、確信度が低い文字はあえて認識結果を表示せず、「＝（ゲタ）」として残し、後工程の人（専門家）に判断を委ねることで、効率のよい翻刻が実現可能な手法である。確信度のしきい値を 0.01 ずつ変化させ、そのときの識別可能な画像数と識別精度を調べたところ、図 16 のようになった。たとえば確信度が 0.63 以上のものに限定した場合、テストデータ 56,029 枚のうち、識別できる文字は全体の 80.98% の 43,401 枚と減ってしまうが、その中で正しく識別できたものは 39,105 枚あり、識別率は 90.10%（=

39,105 / 43,401) まで向上する。この場合、識別できない文字が 22.54% あるが、これらの多くは、

- 学習データに含まれない文字
- 学習データとテストデータで環境が異なり、作者によって文字の書き方が違うことから識別が困難な文字

であることから、この結果は妥当と考えられる。このとき、学習に使われていない（出力結果を表示したくない）文字クラスの画像 3,003 枚の結果を確認すると、80.66% にあたる 2,422 枚が期待通りに排除できることが確認できた。

5. ま と め

本研究では、古典文学の翻刻の作業効率化のため、ディープラーニングを用いて、1,500 クラスを超える大規模なくずし字の認識を試みた。画像数が少ないクラスの画像をデータ拡張した場合、1,506 クラスの文字の識別率は 76.64%、各クラスの平均認識率は 66.00% となった。この識別精度では、実際の翻刻の効率化には不十分なため、誤りの可能性がある文字の画像や、学習に用いられていない文字の画像を自動的に排除する手法を検討した。確信度に注目することで、これらの文字を自動的に排除することを試み、確率のしきい値を変えることにより 90% 以上の識別率を達成できることが確認できた。

しかしながら、歴史的文献の翻刻、電子化の推進のためには、さらに識別精度を改善していく必要がある。精度精度のため、識別率が悪いクラスを分析結果をもとに、以下の対策を検討していく予定である：

- (1) 単純にディープラーニングなどの機械学習技術の改良を図るだけでなく、くずし字に関する専門知識を踏まえた上で改善法を検討する。

- (2) 文献³⁾で行われている大津の二値化を用いたノイズ除去など入力する画像の変更や、ニューラルネットワークのハイパーパラメータの変更など、モデルの最適化を行う。
- (3) データ数が少ないクラスが多いことから、ディープラーニングに加え、文献⁵⁾で提案されているような、文字を書く過程に着目したワンショット学習の技術を取り入れる。
- (4) 同じ文字コードの画像を字母ごとに分ける、時代ごとにデータを分けるなど、文字認識に必要な古文書文字認識用辞書を構築する。
- (5) 文字単位で別々に認識をするだけでなく、前後の文字や文脈から判断し、単語レベル／フレーズレベル／文レベルで認識する。

参 考 文 献

- 1) 早坂 太一, 大野 互, 加藤 弓枝, 山本 和明, “ディープラーニングによる変体仮名の翻刻および WWW アプリケーション,” 情報処理学会シンポジウム論文集, (2016) 7.
- 2) 上田 佳祐, 蘭頭 元春, 飯山 将晃, “CNN と文字のアスペクト比を用いたくずし字認識,” ELCAS Journal, 3, (2018) 88.
- 3) H. T. Nguyen, N. T. Ly, K. C. Nguyen, C. T. Nguyen, M. Nakagawa, “Attempts to recognize anomalously deformed Kana in Japanese historical documents,” International Workshop on Historical Document Imaging and Processing (HIP 2017), (2017).
- 4) Z. Yang, K. Doman, M. Yamada, Y. Mekada, “Character recognition of modern Japanese official documents using CNN for imbalanced learning data,” Proc. of 2019 Int. Workshop on Advanced Image Technology (IWAIT), 74, (2019).
- 5) B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” *bf* 350, 6266, (2015) 1332.
- 6) 山本 純子, 大澤 留次郎, “古典籍翻刻の省力化：くずし字を含む新方式 OCR 技術の開発,” 情報管理, 58, 11, (2016) 819.
- 7) T. Clanuwat, M. B.-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, “Deep Learning for Classical Japanese Literature,” arXiv:1812.01718, (2018).