# 資料と公共性 ： 2018年度研究成果年次報告書

岡崎, 敦
九州大学大学院人文科学研究院 | 九州大学大学院統合新領域学府：教授

市澤, 哲
神戸大学大学院人文科学研究科：教授

石田, 栄美
九州大学附属図書館 | 九州大学大学院統合新領域学府：准教授

後小路, 雅弘
九州大学大学院人文科学研究院 ： 教授

他

https://doi.org/10.15017/2230688

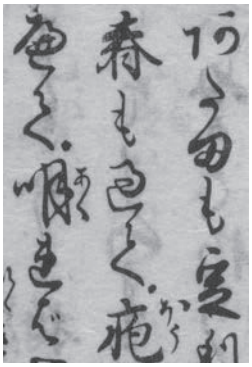
出版情報：2019-03-14. 九州大学大学院人文科学研究院
バージョン：
権利関係：

# くずし字のオープンデータとその活用

畑埜　晃平

近年、人文学と情報学の境界領域であるデジタル・ヒューマニティーズ（DH）研究が盛んになっている。DH 研究のための国際会議 DH（米主導）、EADH(ヨーロッパ主導)、また日本初の国際会議 JADH なども設立され、文国内外から、文学だけでなく情報系といった様々なバックグラウンドの研究者達が活発な議論を交わしている。日本においても、歴史的転籍ネットワーク事業や、人文学オープンデータ共同研究センター（CODH）の設立など、DH 研究の機運が高まっている。

日本国内の DH 研究において、古典籍の自動認識は必要不可欠な技術といえる。日本の古典籍は主にくずし字体で書かれている。くずし字とは、楷書の点画を省略した手書き文字と、手書き文字をもとにした版本の文字を指す（図 1、日本古典籍字形データセットより）。



（図 1）

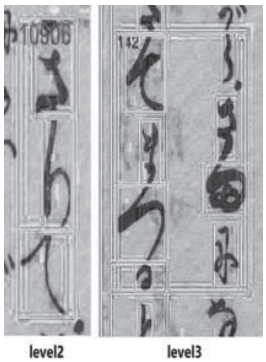
より正確には、くずし字の（自動）認識問題とはくずし字を含む画像が入力として与えられたとき、対応するテキスト（翻刻）を出力する問題である。実は、くずし字認識においては、文字列を文 1 字ずつに区切る事が一番難しい。これはくずし字表記では文字を続けて書くため、文字の区切り情報が含まれていないためである。意外かもしれないが、一旦正しく文字を区切ることができれば、1 文字ずつの認識は現在の AI 技術を用いて高い精度で可能である。

我々は、くずし字の自動認識における、文字区切りの困難性を克服するため、機械学習のアプローチを取る。機械学習とは情報科学の分野の 1 つであり、与えられたデータから未知のデータに関して精度の高い予測を行うための方法論／基盤技術である。具体的には、計算機に文字区切りと文字認識を同時に学習させるのである。そのためには、文字区切り

情報とテキストの情報を含む多量のデータが学習用データとして必要となる。そこで、我々はくずし字のオープンデータを利用し、学習用データセットを構築することにした。

くずし字に関するオープンデータとして、CODH が公開している「日本古典籍データセット」が挙げられる。このデータセットは日本の古典籍 3126 点（2019.1 現在）の画像データ（約 60 万）および書誌データ、テキスト（一部）からなる。また、機械学習用のデータとして、同じく CODH で公開されている「日本古典籍字形データセット」も挙げられる。このデータセットは日本古典籍データセット 28 点から得たくずし字 4645 文字種の字形データを約 68 万文字から構成される（2019.1 現在）。どちらも CC-BY-SA ライセンスの下で公開されている。CC-BY-SA ライセンスでは、データは自由に利用できるだけでなく、２次加工したデータを同じ条件のライセンスで公開してもよい、とされる。これらのデータセットは非常に有用であるが、計算機による学習用データセットとしてそのまま用いることは難しい。というのも、人間と同様に、計算機に対しても、適度な難しさを持つ問題例が必要となるが、当該データセットにはそのような問題例はなく、利用者側で作成しなければならないからである。

一方、2017 年に画像認識コミュニティ PRMU でくずし字認識コンテストが開催され、問題例のデータセットが公開されている[i]。このコンテストでは CODH の日本古典籍字形データセットからコンテスト用の３種類の問題例（LV1：単一文字、LV2：3 文字、LV3：複数文字）を作成している。我々は LV2、LV3 の問題例が全て LV1 の問題例と文字を共有していることに着目した。したがって、LV 1 の 1 文字の区切り情報を LV2、LV3 の 3 文字、多数文字の区切り情報と組み合わせることにより、文字区切り情報を含んだ複数文字の問題例が構成出来る。このようにして我々は文字区切り情報つきの学習用データセットを作成した（図 2）。




（図 2）

今後このデータセットは九州大学附属図書館の機関レポジトリにてオープンデータとして公開する予定である。なお、本成果は国際会議 JADH2018 に受理された[1]。

我々は、作成したくずし字学習用データセットを、くずし字認識手法の学習に実際に活用し、予備実験では既にくずし字認識コンテスト優勝チームの報告結果[2]よりもはるかに良好な認識結果を得ている。優勝チームの既存手法と我々の手法の主な違いは、前者が予め文字区切りの候補を多数作成し、学習手法に与えるのに対して、我々の手法は文字区切りそのものをデータから学習していることにある。

我々のオープンデータ活用例は、オープンデータは異分野からの参入を促進し、新たなアイデアの創生に貢献することを示している。しかし，異分野の研究者にとっては単にデータが公開されているだけでなく、そのデータが持つ意味や意義を知る必要がある。より具体的にはオープンデータだけでなく、解くべき問題は何かを公開すること、すなわちオープンプロブレムが重要と思われる。

**註**

i) https://sites.google.com/view/alcon2017prmu/

**参考文献**

1. T. Yiping, K. Hatano, E. Ishita, T. Nakatoh, and T. Kawahira, “Construction of Japanese Historical Hand-Written Characters Segmentation Data from the CODH Data Sets,” Proceedings of the 8th Conference of Japanese Association for Digital Humanities (JADH’18), pp. 183-185, 2018.
2. H. T. Nguyen, N. T. Ly, K. C. Nguyen, C. T. Nguyen, and M. Nakagawa, “Attempts to recognize anomalously deformed Kana in Japanese historical documents,” Proceedings of the 4th International Workshop on Historical Document Imaging and Processing(HIP2017), pp. 31-36, 2017.