プレスリリース

令和元年７月１０日

「くずし字」の認識に世界のAI研究者・技術者が挑戦
―全世界的コンペティションをKaggleで７月から開催―

日本は、古典籍、古文書、古記録などの過去の資料（史料）を千年以上も大切に受け継いでおり、数億点規模という世界でも稀なほど大量の資料が現存しています。日本の歴史・文化の研究や、過去の災害などの自然現象の解明を進めるには、これらの資料をデジタル化・オープン化するとともに、その内容を読み解く必要があります。ところが、現代のほとんどの日本人は「くずし字」で書かれた過去の資料を読めなくなっており、大量のくずし字をどう読み解くかが重要な課題となっています。

そこでこの社会課題の解決にAI（人工知能）を活用する方法を探るため、この７月から10月にかけて、世界最大規模の機械学習コンペプラットフォームである「Kaggle（カグル）」で、「くずし字認識：千年に及ぶ日本の文字文化への扉を開く」と題する全世界的なコンペを開催します。コンペを通して画期的なくずし字認識手法の開発が進むだけでなく、くずし字データセットを通して日本文化への関心が世界的に高まる効果も期待できます。

本コンペは、情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータ共同利用センター（センター長：北本朝展、以下、CODH）ならびに同機構国立情報学研究所（所長：喜連川優、以下、NII）、人間文化研究機構 国文学研究資料館（館長：ロバート・キャンベル、以下、国文研）が主催します。

日本は、古典籍、古文書、古記録などの過去の資料（史料）を千年以上も大切に受け継いでおり、数億点規模という世界でも稀なほど大量の資料が現存しています。日本の歴史・文化の研究や、過去の災害などの自然現象の解明を進めるには、これらの資料をデジタル化・オープン化するとともに、その内容を読み解く必要があります。ところが、現代のほとんどの日本人は「くずし字」で書かれた過去の資料を読めなくなっており、大量のくずし字をどう読み解くかが重要な課題となっています。

現在、くずし字をきちんと読める人は全国で数千人程度と推定されており[1]、これらの人々だけで膨大な資料を翻刻[2]するには限界があります。この課題を解決するために、２つの方向で研究が進められてきました。第一が市民参加型翻刻システム[3]の開発です。専門家と市民が共に参加する翻刻システムを使い、市民がくずし字を翻刻しながらスキルを向上させることで、くずし字を読める人々の数をもっと増やすことを目指します。第二がコンピ

ュータ（機械）の活用です。機械が文字を読み取る光学的文字認識（OCR）の活用による翻刻の自動化には、これまでいくつもの研究グループが取り組んできました。しかしくずし字は文字の種類が多く、連続した手書き文字の分割が難しく、レイアウトが多様で、本ごとにスタイルが異なるため、実用レベルのくずし字 OCR の研究開発は難航しています。

一方、画像解析の分野における深層学習（機械学習）の活用を中心とした、近年の AI の飛躍的な発展を取り入れることで、新方式のくずし字 OCR に向けた研究開発が進む可能性も高まっています。そこでくずし字 OCR の性能向上に向けたアイデアをオープンに募集するため、CODH、NII、国文研は、この 7 月から 10 月にかけて、世界最大規模の機械学習コンペプラットフォームである「Kaggle（カグル）」[4]で、「くずし字認識：千年に及ぶ日本の文字文化への扉を開く（Kuzushiji Character Recognition: Opening the Door to A Thousand Years of Japanese Literate Culture）」[5]と題するコンペを開催します（図 1）。

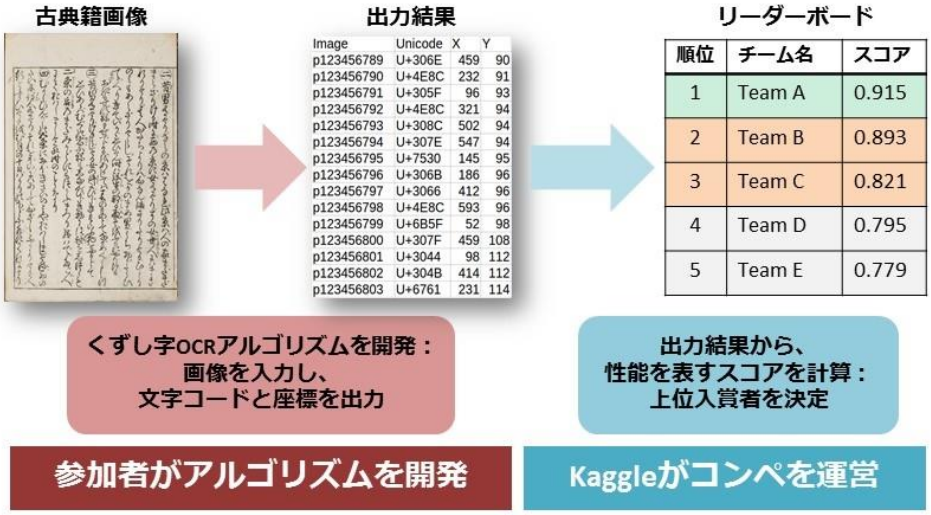

図 1　Kaggle コンペティションの流れ

このコンペでは、国文研が CODH と協力して整備し公開中の「くずし字データセット」をコンペ用に改良して提供します。参加者は、3 ヵ月というコンペ期間内に、与えられた画像内に書かれたくずし字をすべて認識して出力する「くずし字 OCR アルゴリズム」を開発します。そして上位に入賞したアルゴリズムは世界中で自由に使えるよう、コンペ後に公開する予定です。

くずし字認識に関するコンペとして、国内学会などが小規模に開催した例[6]はありますが、今回は世界的な規模のコンペを開催するため、情報学の分野で世界的に知名度が高く、全世界 300 万人以上の AI 研究者・技術者が参加する Kaggle のプラットフォームを活用します。日本の組織による Kaggle コンペの開催はリクルート、メルカリに次ぐ 3 例目ですが、研究目的での開催は今回が初めてです。また Kaggle の歴史の中でも、人文系データを対象とす

るコンペは今回が初めての開催となります。Kaggle 社の関係者も「これまで開催された 337 件のコンペと比べても、本コンペは新しい領域を開拓するものであり、急速に進化するコンピュータビジョン技術の人気を踏まえれば、Kaggle コミュニティが興奮するようなコンペとなるだろう」とコメントしています。

コンペを通して画期的なくずし字認識アルゴリズムが見出せれば、AI による翻刻支援や、AI 文字認識を活用した全文検索など、過去の日本文化を読み解く新技術の研究開発が活発化することが期待できます。そして専門家の作業の一部を AI が支援できれば、専門家は資料の高度な読み解きに集中しやすくなります。このような新技術は、過去の資料を大規模にデジタル化・オープン化し、それを機械や市民が大規模に翻刻し、文理の研究者が過去の世界に関するデータを分析し、その成果を社会に還元するという、データ駆動型[7]の日本文化研究を進めていく上で不可欠になると考えられます。

コンペに関する詳細な情報は、Kaggle ウェブサイト上でコンペ開始日（7 月中旬予定）から公開されます。参加者は、3 か月後の 10 月に設定される締め切りまでにアルゴリズムを提出します。その後、主催者は Kaggle と協力して入賞者（5 位まで）を決定し、11 月 11 日に東京で開催するシンポジウム「日本文化と AI」で表彰式を行う予定です。

【参考資料】

北本 朝展、 カラーヌワット タリン、 宮崎 智、 山本 和明、 "文字データの分析 - 機械学習によるくずし字認識の可能性とそのインパクト -"、 電子情報通信学会誌、Vol. 102、No. 6、 pp. 563-568、 2019 年 6 月、 http://doi.org/10.20676/00000349

【本件に関する問い合わせ】

大学共同利用機関法人 情報・システム研究機構
本部広報室
TEL:03-6402-6214 E-mail：koho@rois.ac.jp

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
総務部企画課 広報チーム
TEL:03-4212-2164 E-mail：media@nii.ac.jp

大学共同利用機関法人 人間文化研究機構　国文学研究資料館
古典籍共同研究事業センター　管理係
TEL:050-5533-2988 E-mail：cijinfo@nijl.ac.jp

[1] 中野三敏 『和本のすすめ』岩波新書 (2011) による。参照：国立情報学研究所プレスリリース「江戸時代の文字の字形データセットを国文研との協働で構築 機械と人間の学習のためのオープンデータとして公開」
(https://www.nii.ac.jp/userimg/press_20161117.pdf)

[2] くずし字の翻刻とは、くずし字を人間が読み、くずし字に対応する現代日本語の文字を入力する作業のこと。

[3] 「みんなで翻刻」(https://honkoku.org/) は、国立歴史民俗博物館の橋本雄太助教を中心に、京都大学古地震研究会や東京大学地震研究所などが協力して構築を進める、市民参加型翻刻システムのこと。CODH も各種の共同研究で協力体制にある。

[4] Kaggle (https://www.kaggle.com/) は、米国に本拠地を置く Kaggle 社 (Google 傘下) が運営する、世界最大規模の機械学習コンペティションプラットフォーム。Kaggle のコンペティションでは、(1)企業や研究者が解決したい課題を出題し関連データを提供、(2)世界中の AI 研究者・技術者がその課題を解決するアルゴリズム (計算手法) を提出、(3) 提出されたアルゴリズムの性能をランキングして上位入賞者を決定、(4) 上位入賞者はコンペの成果を出題者に提供し賞金を獲得、という流れで研究開発をオープンに進める。

[5] Kaggle コンペに関する詳細情報については、下記のサイトで提供する。

本コンペのページ (https://www.kaggle.com/c/kuzushiji-recognition)
※コンペ開始日に公開予定

CODH のウェブサイト (http://codh.rois.ac.jp/competition/kaggle/)

[6] 第 23 回 PRMU アルゴリズムコンテスト くずし字認識チャレンジ 2019 (https://sites.google.com/view/alcon2019) は、2019 年 5 月 31 日から 8 月 31 日まで、電子情報通信学会パターン認識・メディア理解研究会が開催 (CODH 後援)。Kaggle コンペティションと同様のデータセットを用いるが、問題の難易度が異なる。

[7] データ駆動型研究とは、機械学習 (AI) による大規模処理なども活用しながら、(ビッグ) データの収集と分析から得られる証拠に基づき、新しい知見や知識を獲得することを目指す研究方法である。