Title	くずし字×AI : オンラインで世界に開く日本古典籍
Sub Title	
Author	カラーヌワット, タリン
Publisher	慶應義塾大学デジタルメディア・コンテンツ統合研究センター
Publication year	2020
Jtitle	慶應義塾大学DMC紀要 (DMC review Keio University). Vol.7, No.1 (2020. 3) ,p.73- 79
JaLC DOI	
Abstract	
Notes	特集 DMC研究センターシンポジウム第9回「大学教育のミライ： オープンエデュケーションのその先へ」これからのMOOCの話をしよう 開催日時 : 2019年11月20日 (水) 14:00〜19:00 開催場所 : 慶應義塾大学日吉キャンパス来往舎2F大会議室 講演2 オンラインで世界に開く日本の文化財
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO32002001-00000007-0073

講演 2

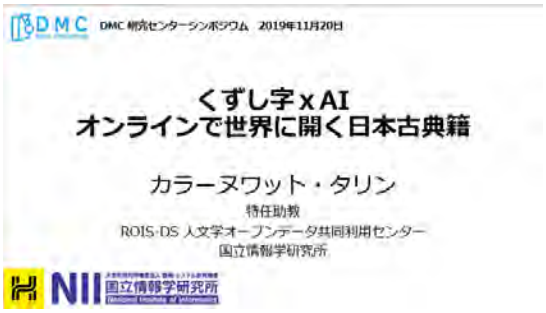
オンラインで世界に開く日本の文化財

くずし字×AI
オンラインで世界に開く日本古典籍

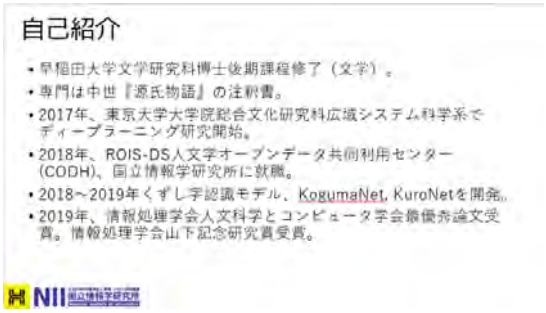
カラーヌワット・タリン

(ROIS-DS 人文学オープンデータ共同利用センター、国立情報学研究所特任助教)

カラーヌワット・タリンと申します。きょうの私の話ですが、キーワードは、くずし字と AI です。AI は人工知能のことなのですが、実は、私は AI を作れば作るほど、この単語が好きではなくなりました。この発表では、AI という単語ではなく、機械学習という単語を使います。



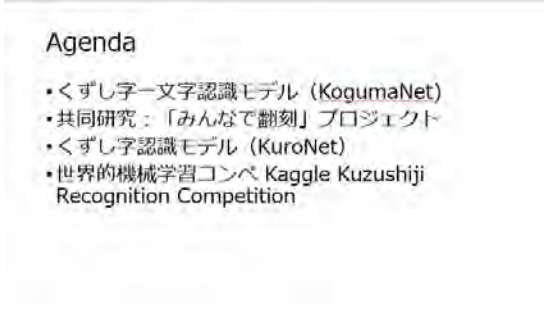
私が MOOC とどう関係しているのかといいますと、私は MOOC で人生が変わった人間の一人なのです。まず、自己紹介をいたしますと、私は 2008 年、タイのバンコクから日本に来ました。そして、早稲田大学でずっと日本古典文学を研究しておりまして、専門は『源氏物語』です。2017 年に、東京大学大学院総合文化研究科広域システム科学系にて、ディープラーニングの研究を開始いたしました。早稲田大学の私の指導教員は『源氏物語』を専門に研究していらっしゃる陣野英則先生でした。2017 年に「私はこれからディープラーニングの研究がしたいです」と陣野先生に言いましたら、先生から「ディープラーニングってなんですか？」と聞かれました。その後、東京大学の山口泰先生の研究室に入りまして、ディープラーニングの研究を開始いたしました。山口先生は、画素処理、イメージプロセッシングの専門で、私がやりたかったのはくずし字です。



私がやりたかったのは、どうすればくずし字を AI で解読できるのかという研究なのです。1 年未満ぐらい、東京大学にいましたが、何をやっていたかというと、MOOC で猛勉強をしました。Coursera や edX など、あるものを何でも使いました。『源氏物語』の研究者から、ディープラーニングをやれる研究者になるためにはどうすればいいかということを考え、基礎知識や専門知識を MOOC で本当に毎日勉強いたしました。そのおかげで、2018 年、人文学オープンデータ共同利用センターに就職できまして、現在、

くずし字認識の研究をメインでやっております。

2018 年から 2019 年に、くずし字認識モデルを開発しました。AI のソフトウェアを私たちはモデルと呼んでいますが、後で説明するように、私たちは KogumaNet と KuroNet という二つのモデルを開発いたしました。また NekoNet という、KuroNet のプロトタイプとなったモデルもあります。私のモデルネーミングは動物を使うことが結構多いです。ただ、Neko と意味なく付けたわけではなくて、Neural End-to-end Kuzushiji OCR の略で Neko としました。それが、情報処理学会人文科学とコンピュータ研究会で最優秀論文賞を受賞し、その後、情報処理学会の山下記念研究賞も受賞しました。



きょうはまず、くずし字一文字認識モデル KogumaNet のことについてお話しします。そして、私と共同研究をやっている、「みんなで翻刻」プロジェクトと、くずし字認識モデルの KuroNet についてご紹介します。KuroNet は、1 ページごとに 1 秒でくずし字を現代日本語文字に置き換える AI のモデルです。そして、KuroNet の研究から発展した、世界的機械学習コンペティション、Kaggle Kuzushiji Recognition Competition についてもお話しいたします。

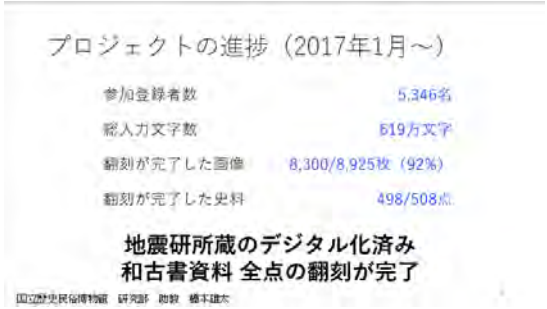
まずは、KogumaNet なのですが、どうして Koguma という名前を付けたのかというと、この AI モデルはものすごく小さくて、モデル自体は、ただの 5 メガバイトなのです。

あまりにも小さいので、これには Koguma という名前を付けようと思ったのです。このように小さいので、携帯でも iPad でも使うことができます。これは何をしているのかというと、古典籍の上に、もし読めない文字があったら、そのバウンディングボックスをマウスや指でいいのですが、それを囲うと、AI がその文字を読んでくれるのです。この文字を読むにあたっては、多分この文字だろうという確率を数字で出しています。

「みんなで翻刻」は、国立歴史民俗博物館助教の橋本雄太さんが開発したものなのですが、これはどういうものかといいますと、先ほど申し上げましたが、日本にはたくさんの古典籍や古文書があるのですが、読める人が非常に少ないのが現状です。

「みんなで翻刻」は、資料をクラウドソーシングして、「みんなで翻刻」を使って、協力しあいながら翻刻するオンライン作業のことです。「みんなで翻刻」のアプローチは二つあって、まず、くずし字解読の学習サービスのフェーズがあり、アプリ経由でくずし字をどう読めばいいかということを学習します。そして、ある程度学習して知識が身についてきたら、その次のフェーズで、自然に、翻刻作業に参加してもらうのです。翻刻は、くずし字の文字を現代日本語に置き換えるという作業です。

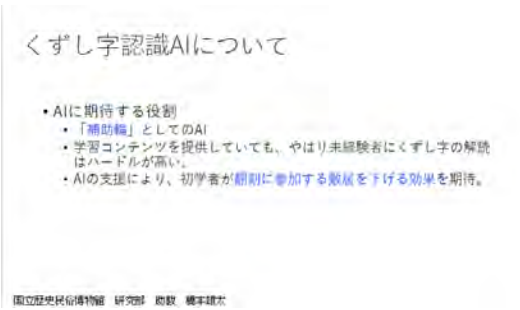
「みんなで翻刻」のユニークなところは、タイムラインがあって、これは SNS のようなものなのですが、例えば、誰が何文字まで翻刻したのかがタイムラインに出ます。そして、もう一つついいところは、添削の機能もありまして、例えば、勉強したばかりの初心者がいきなり翻刻作業に参加するというのは、非常にハードルが高いのですが、そういう方には上級者の方が添削してくれるのです。翻刻した結果を正しくするのがこの機能なのです。



このプロジェクトは 2017 年から開始し

て、参加人数は 5000 人以上になっています。翻刻した文字数は 600 万文字。東大地震研の資料から始めたのですが、最初に対象とした資料は全て翻刻が済んでしまいましたので、今後はいろいろな古典籍にまで拡大しようとしているところです。

「みんなで翻刻」のくずし字認識 AI には、先ほどご説明した KogumaNet を取り入れています。AI の役割は自転車の補助輪のようなものです。やはり未経験者にはくずし字解読は非常に難しく、ハードルが高いのですが、AI を使うことで、AI がこの文字は何かということを推測してくれます。ただし、AI は、100 パーセント正しいわけではありません。私はくずし字を読めますが、AI が読んでくれたものと、私が読んだ結果が違っていて、AI のほうが正しくないという結果のときでも、言われてみれば確かにこの文字に似ているなと思う経験が何度もあります。この AI の支援により、ハードルをできるだけ低くして、初心者が参加しやすくなるようにします。このように KogumaNet は活躍しております。



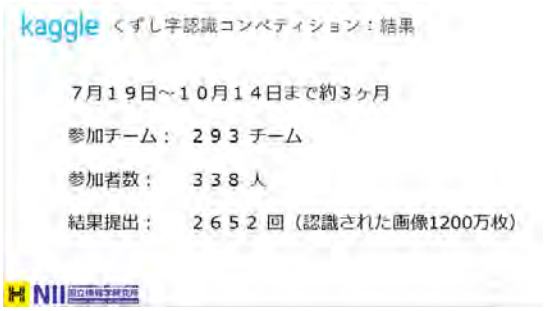
次はKuroNet です。KuroNet というAI は、1 ページごとにモデルに入れて、1 秒で翻刻の結果が出ます。この本は『徒然草』ですが、文字は全て 1 ページごとです。このモデルを制作したときの論文もありますので、もしアルゴリズムについて興味がありましたら、論文を読んでいただければと思います。

KuroNet は、何ができるのかというと、例えば、国会図書館のこのからくりの本をみるとき、くずし字を読めない人がどう見るのかと考えたとき、そういう方は、おそらく、絵を見ているのです。そこで、この絵はどういうものなのかということを解読しなければならないのですが、KuroNet を使うと自動的に文字を置き換えることができます。これは『道成寺、鐘うつからくり』という文字が出てきます。このように、くずし字の教育がない人でも、古典籍を読めるようにするという役割を担っています。

次に、KuroNet をベースにして、7 月 19 日から 10 月 14 日まで、Kaggle Kuzushiji Recognition Competition を開催しました。これは機械学習コンペティションです。くずし字認識に関するコンペとして、世界中から機械学習研究者が参加できるコンペティションを開催しました。

このキーワードの Kaggle とは何かといいますと、これは会社名でして、現在は Google の子会社になっています。Kaggle は、機械学習の研究者が集まる、世界最大のデータサイエンティストコミュニティーで、すでに登録者数は 300 万人を超えています。AI 機械学習やデータサイエンティストなど、登録者はいろいろな方がいます。Kaggle コンペティションの非常にいいところは、参加者の国籍、年齢、技術レベルを問わず、世界中から誰でも参加できる仕組みになっていることです。ただ、未成年者は、保護者の許可が必要です。

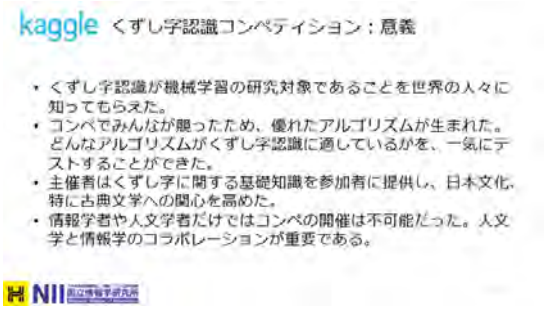
このくずし字コンペティションは、読むときに専門知識が必要なのではないかということが課題になったのですが、このコンペティションをデザインしたときに、日本語ができなくても、くずし字が読めなくても、参加できるようにデザインをいたしました。このコンペを 3 カ月間開催して、参加者は 300 人以上となりました。また結果を 1 回提出するために、くずし字のページを 4551 枚認識しなければなりませんが、全部で 2652 回の提出がありましたので、合計で 1200 万枚の画像が認識されたことになります。



先ほどお話しいたしましたように、このコンペティションは日本語が分からなくても、くずし字の知識がなくても、誰でも参加できるということにいたしましたので、残念ながら、1 位は日本ではなくて中国でした。F1 スコアという数字から見ると、95 パーセントぐらいの精度でした。2 位はロシアです。そしてようやく 3 位に日本が入りました。ドイツ、ロシア、中国など本当にたくさんの国の方が参加してくれました。

まず、どうしてこのコンペティションをやろうと思ったのかというと、私たちの開発したくずし字認識のアルゴリズムを、より良いものにしたいと思い、コンペを開催したということがあります。もう一つの目的は、くずし字認識は、機械学習とあまり関係がないように見えるのですが、この認識問題のことを機械学習の研究問題として世界中の人々に知ってもらいたいという気持ちもありました。

そして、Kaggle のコンペティション開催の意義は何かといいますと、コンペティションでみんなが争ったために、優れたアルゴリズムが生まれたことです。私たち自身は、KuroNet という一つのアルゴリズムだけしか試すことができませんでした。その理由は、たくさんの時間や努力が必要だったからなのですが、それをコンペティション経由で、300 人ぐらいの参加者が、みなさん違うアルゴリズムを作ってくれましたので、どのような手法が一番くずし字に適しているのかということを、一気にテストすることができました。また、コンペティションを開催するときに、参加者の皆さんには少なくとも、くずし字とは何かということを、基礎知識として提供いたしました。例えば、どうしてくずし字があるのか、どうして現代の日本人はあまり読めなくなってしまったのかなど、そういう歴史的な知識を提供して、日本文化、特に古典文学への関心を高められるようにいたしました。

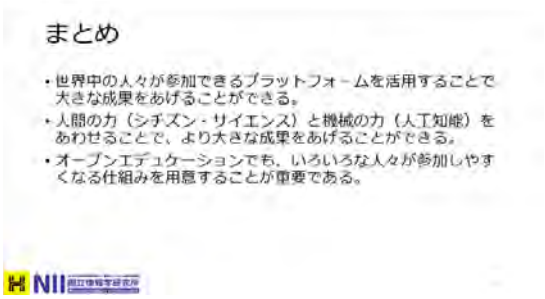


Kaggle のディスカッションで、どうして

この文字はこう読むのですかというような質問や、どうして同じ平仮名なのに、いろいろな形があるのですかという変体仮名などについて、質問をたくさんもらいました。皆さん機械学習研究者で、特に海外の機械学習研究者は、日本の古典籍とはほぼ無縁なので、このコンペを経由して、古典籍とは 1 ページでこういうものなのだ、あるいはこういう文字を書いているのだというような部分を見てもらえました。

そして、もう一つ大事なことは、このコンペティションは情報学者や人文学者だけでは、開催することができなかったということです。これは人文学と情報学のコラボレーションから生まれたコンペティションなので、今後、違う分野の人たちが、協力して何かを作っていくことで、非常に素晴らしいものが作れるのではないかと思いました。

そして、まとめといたしましては、世界中の人々が参加できるプラットフォームを活用することで、大きな成果を上げることができるということです。そして、もう一つ AI のことなのですが、AI は職を奪うものと考えている人がたくさんおります。でも、私は、人間の力と機械の力を合わせて、より大きな成果を上げることができるのではないかと考えました。そして、オープンエデュケーションで、ハードルを低くすることによって、誰でも参加できるようにすることが非常に大事なことだと思っています。これによって、さらに良いものを作れるようになるのではないかと思います。



少々早口になりましたが、私の発表はこれで終わりにいたします。ご清聴ありがとうございました。