AI 活動による古典研究の可能性

同志社大学文化情報学部助教　深　川　大　路

はじめに

文化情報学部の深川でございます。「AI 活動による古典研究の可能性」というタイトルでご報告させていただきます。自己紹介と、授業でどんなことをやっているのか、「くずし字翻刻」についての研究の話、終わりに今後の展望をお話させていただきます。私は計算機科学、離散最適化、アルゴリズムということを研究している者で、コンピュータで効率よく計算するにはどうしたらいいか、コンピュータをうまく使うにはどうしたらいいか、コンピュータを使っても、効率よく計算しようとしても難しさがある、複雑な問題があり、それを数学的に解析しようということを、これまでやってきております。ホームページにも書いていますのでご興味のある方はそちらを参照してください。

私と司会を務めております福田さんは文化情報学部に所属しています。京田辺校地の夢告館にありますが、データサイエンスで文化を探究する文理融合型の学部です。文化とは何か、芸術から経済活動など社会の動きまで人間の営みすべてを「文化」と捉えようというのが文化情報学部のスタンスです。その「文化」はこれまで自然科学の対象とはなりにくかったわけですが、それをデータサイエンスの手法で探究するという比較的新しい試みをしている学部です。詳しくはホームページ等をごらんください。

1　古典籍を対象とする授業運営と情報システム

文化情報学部では「ジョイント・リサーチ」に取り組んでいます。「ジョイント・リサーチ」は文化情報学部３年生の必修科目で、この特徴は何か。グループ単位で研究活動を行う。教室に座って先生の話を聞いて「ふむふむ」というようなことではなく、自ら創意工夫を凝らして研究を行う趣旨の科目です。「ジョイント・リサーチ」はクラスごとに分かれて受講しますが、「ジョイント・リサーチ」のクラスを福田さんと私で担当しています。二人で担当しているといいましても、いろんな先生方にご協力を賜りまして長く続いているクラスです。途中、タイトルは変わっていますが、これからお話するようなことと似たことの研究を続けて取り組んでおります。

「情報科学技術を用いた日本古典文学の文字列解析」。まさに文理融合かなという内容です。「伊勢物語御歌かるた」は京田辺校地のラーネッド記念図書館に所蔵されています。「春日のゝ、わかむらさきの　すり衣　しのぶのみだれ　かぎりしられず」。有名な歌ですが、このかるた、百人一首かるたと違いまして、『伊勢物語』の物語の、ある場面が書かれていて、きれいなかるたです。

「伊勢物語歌かるた」を史料として、それがどのような特徴をもっているか。「絵札」と「字札」がある。百人一首かるたと同じですが、絵札の裏には下の句が書かれています。どのようにこのかるたが使われているかは、よくわかっていないという話を聞

いた覚えがあります。競技かるたのような使われ方をしていたのかどうか、わかりませんので、「読み札」という言い方は避けて、「絵札」と「字札」と、ここでは呼んでおります。

「絵札」の表に、きれいな模様が書かれていて裏に下の句が書かれていて対になっているものです。もう一つの特徴は現代のかるたとは違いまして「くずし字」ですが、特に、特徴は「散らし書き」で、「みだれ初にし」と書かれています。「我」「ならなくに」とあります。「みだれ初にし　我　ならなくに」という書かれ方が整然とは並んでいない。原稿用紙に書くような順番で並んでいない。コンピュータが、人間が読む、人間が知識もなく読むのは、かなり難しいのではないかという特徴があります。それをあえてコンピュータで解析をする。ここでは「翻刻」をする。字を読むということですが、つまり文字が書かれた画像の上に文字を重ねて、その文字を現代の印刷物のような読みやすい字にできたら多くの人に使っていただけるのではないかと。

コンピュータではなくて、紙の上で翻刻するにはどうするか。書き込んだりすることができるわけです。ペンと紙があればいいわけです。かるたの画像をプリンタで印刷したものに書き込む。それをコンピュータ上でできたらいいなということで「画像注釈支援システム」をつくり、授業でも使っていました。ウェブサーバを夢告館に設置して、その上にシステムをつくり、このような画面にしております。ここには男女が描かれている。ここには文字が書かれている、そういうものをデータベース化できるシステムをつくってみたりしました。それを工夫しまして、注釈を更新

しやすいように、授業の中で扱いやすいようにするにはどうしたらいいか。たとえば学生が操作になれているようなエクセルデータを格納するファイルとして採用し、エクセルファイルを修正してサーバに戻すことで「注釈」が置き換わる仕組みをつくったりしました。簡単にいうと、かるたの画像がデータベースの中に格納されていて、それともう一つ、その画像にどんな文字が書かれているかという注釈データベースと、その二つのデータベースをつなぐシステム、そのシステムにユーザがアクセスして字を読んで、データをつくったり、閲覧したりするシステムです。これを、USB を用いてオフラインで使えるように改良したり、いろんなことをしました。ウェブ上で使えるようになったシステムを、さらに変えて文字ごとにずらずらと並べたり、「字母」ごとに細かい改良をする。可能・不可能の「可」というのと加えるという「加」の 2 種類の漢字を成り立ちとするもの。少し形が違う。上の 3 つは可能の「可」をもとにして一番下は「加」をもとにして少しグループが違うことが見てとれるかと思いますが、ひらがなの元になるもの、「字母」といいますが、それがどれくらいの割合でデータの中に入っているかを表示したり、一つの「字母」をもとにした文字だけを列挙し、表示させるシステムをつくったりしておりました。

「挿絵」についても同様で「どんな挿絵が、どれくらいあるか」を閲覧できるように工夫しておりました。お遊びですが、せっかくデータができたので何かできないか、と。学生さんが「変体仮名」を学ぶ時、楽しく学べないか。ゲームをつくってみようと。

これは真ん中にある文字をコンピュータのマウスでズルズルと動かしていきますと、ここに重ねてカチッと嵌まると「カチッ」という音を鳴らしたい。もうちょっとできたら派手な音を鳴らすとか楽しげなゲームをつくってみたいと。データベースを、より閲覧性、操作性を高めるために、つくり替えたりしてきております。Django というシステムを使っています。興味がある方はこちらの URL をご覧ください。（https://tiramis2.doshisha.ac.jp/db/）

昨今、画像の史料をどのようにすれば、効率よく、共有してブラウザで閲覧できるかということでつくられた国際規格 IIIF があります。それに則ったシステムにつくり変えていくと、いろんな画像を閲覧するためのソフトがありまして、IIIF に対応した、さまざまなビューワーで閲覧者が好きなビューワーを選んで同じ一つの資料を閲覧できるというメリットが得られるということがあったりします。

文字をズラズラと並べて「字母」ごとに表示したりするシステムの一つの形ですが、このシステムの裏側はどうなっているか。グーグルがスプレッドシート、エクセルの表をクラウド上でウェブブラウザだけで利用できる便利なサービスがありまして、学生さんが複数名で、グループで活動する時に一つのコンピュータに、みんなが集まらなくても作業ができるというメリットがあります。共同で、オンラインで自宅からも大学の教室からもアクセスができて、一つの作業をグループで行い、一つのかるたの翻刻を行うシステムをつくっていました。それがグーグルのスプレッドシートを使うとリアルタイムでデータベースに反映する仕組みを

簡単につくることができるということで採用しました。ということが授業で、どんなことをやっているかというご紹介でした。

2　くずし字翻刻支援システム

これまでは人手で翻刻データをつくっていました。「くずし字翻刻」とはそもそも具体的に何をしているか。日本古典籍における「くずし字翻刻」というのは画像から文字を抽出して文字の識別を行う。具体的には画像から文字を抽出してデータ化する。翻訳のように意味を読み取ることまではしない。まだ字を読むだけが目的となっています。文字順序とか位置を気にする場合もあります。

コンピュータでは情報を数字に変換して扱います。文字もコンピュータで扱う時には数字に変換します。その時に使われるのが「文字コード」で、「あ」だったら数字に置き換える。スライドには３つの数字を挙げています。なぜ３つになるのか。文字を数字に置き換えるための方式が、さまざまあるからです。「文字コード」にもいろんな種類があります。現在は Unicode が広く用いられています。その「文字コード」の決まりに従うことで、いろんな計算機が世界中にあります。スマートフォン、コンピュータ、みんながもっているもので、お互いに情報のやりとりができる仕組みになっています。「くずし字翻刻」も文字を数字に置き換えることが一つの目的になります。

「くずし字翻刻」には二つの方法がありまして、まず手書きで

やる方法。古典籍は手書きのものが多い。活字もありますが。手書きのものを読み取ることはコンピュータからすると、同じ文字が全く同じ形で書かれることがほとんどないという難しさがあります。もう一つは連綿体、つづけ字で書かれていまして、一つの字がどこまでなのかが、わかりづらい。文字の切れ目がわかりづらい。もう一つは「変体仮名」で文字が書かれているケースが多い。現代の日本語は、ひらがな、カタカナ、漢字がありますが、近世以前、明治初期までは「変体仮名」が広く用いられております。珍しいものではありません。それに対応する必要があります。「変体仮名」は「あ」ですと現代のひらがなの「あ」、安心の安をくずしてつくられた仮名ですが、それではなくて、「阿」、「悪」などからつくられたものも「あ」として相互に交換可能な形で使われる。「変体仮名」をコンピュータで扱う場合、文字を数字に置き換える必要がありますが、その取り決めがなかったんです、最近まで。2017 年に、ようやく Unicode に「変体仮名」が追加されることになりまして、それでコンピュータで「変体仮名」を扱うことがしやすくなった。しかし Window をそのまま使うと「変体仮名」ができないことが多い。フォント、文字を表示するための仕組みが対応していない。まだまだ整備が必要ですが、整いつつあります。

「くずし字翻刻」というのは、そもそも何のためにやるのか。テキストを理解するための第一歩。古典籍があり、史料がある。しかし読めないことには内容がわからない。古典籍のテキストを理解するのはどういう目的があるか。日本の伝統文化の記録を読

み取ることは深い意義があるわけです。

「くずし字翻刻」の目的の二つ目は「機械可読なデータの作成」。機械が読めるデータ、数字に置き換えてある。言い換えるとデジタルデータという言い方になります。デジタルデータに置き換えること、数字で置き換えることによって何ができるか。統計処理、一括処理ができるようになり、たとえば大規模な文字を数えたりできます。数えるという作業は単純ですが、量が多かったり繰り返し行ったりする場合、人間だと面倒になったり、間違えたりするわけですが、コンピュータだと同じ調子で作業を続けてくれる。もう一つは「デジタルアーカイブや複製の作成」も「くずし字翻刻」の目的です。現存する貴重書の経年劣化は避けられないこともあります。後世に伝えるための方法の一つとして複製をつくる。富士ゼロックスの試みも、そうですが、現代の一つの写本の形としてデジタルアーカイブがあるのではないか。実物を実際に使うことに対しても、伝えるだけではなく、活用することも大事で、触ることができる。コンピュータで貴重書を閲覧しても現物が壊れることがないのでリスクがない。

「くずし字翻刻」の現状について。多くの有名な古典籍、『源氏物語』『伊勢物語』とか誰もが知る古典籍は誰かの手で、すでに翻刻されています。伝本ごとにも翻刻されています。一方、まだ翻刻されていない古文書が世の中に大量にあるだろうと、すぐに想像がつくところかと思います。どれくらいあるか、正確な把握は難しいのですが、国文学研究資料館の歴史的典籍 NW 事業では毎年 38000 点の歴史的典籍を撮影するプロジェクトが継続してい

ます。その結果、30万点の画像公開を目指して現在、22万6千点が、すでに撮影済となっています。画像コマ数でいうと2400万コマ、専門家だけの手では難しいため、市民参加型の翻刻プロジェクトがあったりします。一般の非専門家と専門家が手を取り合って翻刻するプロジェクトもありますが、かなり成功している部類にはなりますが、2400万コマはなかなか難しい。ではどうするか。コンピュータの力を、うまく使うのがいいのではないかと。人間の手のみによる翻刻には限界があります。翻刻に携わる専門家や市民の数を倍にしようといっても一朝一夕にはいかないわけですが、コンピュータ、AIの数を倍にすることは簡単にできます。最近、半導体の値上がりとかもありますが、人を倍にするより、ずっと簡単なわけですね。

AIを用いた自動翻刻も、そんなに精度は十分ではないということがありますが、凸版印刷の報告によれば、90％の精度のあるOCRが実現したという報告がありますが、間違いはつきものです。人間も間違いますが、コンピュータの方が人間の専門家に比べれば、まだまだ劣っているところがあります。発展途上の段階なので、これからはわかりませんが、現代はそういう状態です。「協働」ということで人間とコンピュータが手を取り合ってやっていくといいのではないかと考えてシステムをいろいろと触っております。

矩形、四角形で文字が囲めないような形で書かれている文字、四角形で枠をつけても、これが「し」ですよ、とデータをつくったら他の文字が混ざったりします。「文字の切り出し」は難しい

問題があります。矩形でなく、不定型な文字を切り出すことができて、コンピュータに「これが「し」ですよ」と教えることができれば、よりうまく扱えるようになるわけですが、それが今の技術では、そう簡単にはいかない。人間と協働しながら文字を切りだしていこうというシステムを「ジョイント・リサーチ」の授業で使ったりしておりました。

「変体仮名」の扱いについては、まだこれからですが、「変体仮名」をどうやってデータベース化していくか。「せをはやみ」は崇徳院の有名な百人一首の歌ですが、「み」という字を3つ並べていますが、「字母」としては真ん中の「み」は現代の「み」で、左右は違っている。「変体仮名」の違いから何か面白い分析できるのではないかという可能性があり、そのためにデータベースを精緻にしていく必要があります。自動的に「変体仮名」の区別をすることができないかと、2019年の終わり頃、学生さんたちががんばりまして、「字母」の違いを考慮することによって文字の認識精度が向上するのではないかという取り組みをして、少し精度の向上がみられたという結果も得られております。

「変体仮名」や「くずし字翻刻」に関する情報システムの取り組みが、いろいろあります。「変体仮名」を学ぶためのアプリを早稲田大学とか大阪大学でつくったものがあったり、海外でもムービー的な解説があったり、自動的に「字を読む」という「深層学習」を用いた「くずし字翻刻」のシステムをつくっているところがあったりします。

「みんなで翻刻」というプロジェクトとCODHという機関がつ

くった「くずし字認識」OCR システムとか、凸版印刷の「ふみのはゼミ」とか「くずし字認識」として、その成果が日本古典籍データセットとして公開されています。公開されたデータを使って AI の精度を高めるという循環ができつつあります。

おわりに

このような流れの中で、今後、どうしていくことが望まれているか。「くずし字データベース」の拡充。画像と翻刻情報の連携方法を標準化することによって、一つの研究室だけではなく、広い開発者と手を取り合い、いろんな機能を追加していくためには、オープンかつ広く利用されている技術を採用してシステムを少し調整していく必要があるだろうと考えています。翻刻作業のユーザインターフェースを高性能化することにより、CODH のシステムと連携するとか、他の「くずし字認識」AI と連携することとか「変体仮名」への対応、「文字の切り出し」が、矩形でないような、複雑なものであっても扱えるようにするとか、そういうことを行っていく。「くずし字データベース」をどんどんつくっていくことも、そうですが、それを活用するところも大事だと考えております。

このようなポータルサイトを用意していますので興味がありましたら、ぜひご覧いただければと思います。以上でございます。