2024.12.24 開催
第 12 回 ACADEMIC CAFE
「AI と人の協働」

AI を用いた古文書解読の最前線

文学研究科　准教授　向井　伸哉

概要　近年、歴史学（中世フランス史）の分野にも、デジタル技術が急速に浸透してきている。AI を用いて手書きの古文書を瞬時に文字起こしする HTR（Handwritten Text Recognition）の研究開発が、目下、史学・工学の分野をまたいで推進されているところである。今回の報告では、HTR の代表的アプリである Transkribus とその技術を利用した Himanis プロジェクト（80000 ページを超える中世フランス国王文書のデジタル化計画）を紹介しつつ、文理融合の最先端の現場をレポートしたい。特に、HTR 登場以前の人力による古文書解読プロセスを振り返ることで、人間による古文書解読と比べた時の AI による古文書解読の長所を確認しつつ、同時にそのダークサイド（短所）についても思いを巡らしてみたい。

キーワード　HTR（Handwritten Text Recognition）、Transkribus、Himanis プロジェクト

座談会の様子

1. HTR（手書きテクスト認識技術）の登場

昨今、デジタル技術が歴史学にも浸透し、手書きの古文書を AI に瞬時に解読（文字起こし）させる HTR=Handwritten Text Recognition（手書きテクスト認識技術）が注目を集めている。活字に対応した OCR（光学文字認識）はよく知られているが、手書きテクストの場合、印刷されたテクストよりも、文字そのものの認識にくわえて、行・レイアウト、縦書き／横書き等の認識が複雑になるため、それらが高い精度で行えるプログラムが求められるのである。

代表的な HTR アプリ Transkribus（https://www.transkribus.org）は、READ（Recognition and Enrichment of Archival Documents）プロジェクトのもと、オーストリアのインスブルック大学を中心に、スペインのバレンシア工科大学、ドイツのグライフスバルト大学、フィンランドのフィンランド国立公文書館など 12 の研究機関によって共同開発されたものである。搭載しているディープラーニングモデル（多層人工ニューラルネットワークモデルによる深層学習を行う文字認識エンジン）は、「PyLaia」や「CITlab HTR+」である。手書き文書の画像 100 ページ分をまず手動で翻刻することが推奨されており、元画像（問題）と手動で入力されたデータ（解答）をグラウンドトゥルース（機械学習の学習材料）として用いて、モデルのトレーニングが行われる。もし、対象の文書に近い言語や書体の公開モデルがサイト上にあれば、それをベースにして、新しい翻刻

データでトレーニングし、独自にカスタマイズしたモデルを作成することも可能である。

HTR が目下直面している問題としては、u と n の区別や略語の判別の困難のほか、機械学習の学習材料の不足が挙げられ、最後の点については、github を通じてグラントゥルースを共有可能にする試みが進んでいる。

2. HTR の応用：HIMANIS プロジェクト

最新の HTR 技術を駆使したプロジェクトとしては、HIMANIS project（2015-2018）がひときわ目を引く。同プロジェクトは、フランス国立科学研究センター（CNRS）のテクスト史研究所（IRHT）が主導し、A2iA（人工知能と画像分析を専門とするフランスの企業）、フローニンゲン大学（オランダ）、バレンシア工科大学の研究チーム「tranSkriptorium」（スペイン）の連携を得て、資金約 40 万ユーロ＝6600 万円を投じて遂行された。14-15 世紀フランス王国の文書局が作成した総量 199 巻、実に 80000 ページ以上に及ぶ「国王文書の宝物庫の登録簿（registres de Trésor des Chartes）」がデジタル化され、付随的な情報（メタデータ）のタグ付けが行われた。複数の書記によって複数言語（ラテン語、古フランス語、オック語など）で書かれた文書群を、全体の 0.5%に満たないグラウンドトゥルース（既存の公刊史料など）を基に、「tranSkriptorium」が独自開発した HTR アプリ（文字認識エンジンは PyLaia）に機械学習させ、手書き文字の判読、略語の判別、書記の特定などを高い精度で実現することに成功している。もちろん、特定のキーワードを指定してコーパス全体の横断検索を行うことも可能である。

実際に HIMANIS のサイト（http://himanis.huma-num.fr）に移動し、古文書画像上の特定の単語にカーソルを合わせてクリックすると、文字起こしされた単語の候補（①super、②supra）とそれぞれの信頼度（①59.60%、②40.40%）が表示される。当然、信頼度が 100%に近い場合は、候補は一つしか示されない。手書き文字認識の精度はたしかに高いものの、間違いも散見され、人間がより多くの学習材料、すなわち古文書画像（問題）とその翻刻（解答）を準備して訓練させる余地がある。くわえて、より正確な解読のためには、文法事項（単語の格変化、複数の単語の慣用的な結びつき・・・）の学習が必須であろう。とはいえ、将来的には、人間よりも正確に解読（文字起こし）するようになることは間違いなく、歴史研究者は、この AI の利用により大幅に時間、費用、体力を節約することができるだろう。

3. AI を用いた古文書解読の短所？

反面、「一単語一単語につまずきながら遅読を強いられることで、新たな気付きや着想につながる」、「解読に際しさまざまな二次文献を参照する度に、新たな研究テーマとの偶発的な出会いのチャンスが生まれる」、「自分の手垢のついた辞典や故人から受け継いだ辞書をめくることで、モチベーションが上がる」、「難読箇所を判読できた時の高揚と喜び」といった人力による解読の長所は失われることになる。古文書解読の AI へのアウトソーシングとは、豊かな魅力や可能性を孕んだ研究プロセスの省略であるともいえる。

一つ例え話をしよう。富士山山頂で、朝日の写真を撮るのを趣味とする人は多い。近い将来、山頂まで一飛びの空飛ぶタクシーが登場したら、彼ら彼女らは、コスパ・タイパを重視して、空飛ぶタクシーを利用し、山登りのプロセスを省略するだろうか？山登りのプロセスも、古文書解読のプロセスも、手段であると同時に目的でもあることを考えると、技術的にアウトソーシング可能になっても、人間がこれを止めることはなく、また、止めるべきではないように思われる。正しい結果を導き出す速度・効率性と引き換えに、研究プロセスの途上で得られる喜びと輝きは失われ、思いもしない新たな研究テーマへと導かれる世界線も消え、長い研究プロセスの末にたどりついた「発見の鋭い喜び」もまた人間の手からこぼれ落ちてしまう・・・。AI の過剰な利用は、AI への「人生のアウトソーシング」に陥る危険性をはらんでいるのではないだろうか。

4. 文学と科学の間で―座談会を終えて―

この度の報告は、古文書解読という歴史学研究の基礎的段階への AI 導入をめぐる話に終始し、解読した古文書を使ってどのように研究するのか、そこに AI が入る余地はあるのかといった、研究の本丸にかかわる問題には立ち入らなかった。他方で、池田先生のご報告は、新たな有機半導体の合成を予測するという研究の核心部分における AI 利用を正面から扱うものであった。ある特定の結果を導き出す因子の種類・数を明らかにすることと、個々の因子が結果に対して及ぼす効果の大きさを明らかにすることは別の作業であり、池田先生が紹介されたデジタル有機合成のケースでは、AI による機械学習（ビックデータ分析）は後者に強みを発揮しているように思われた。そうだとすると、分野を問わず、前者に関する AI の活用事例としてはどのようなものがあるのか興味をそそられた。

また、監視カメラの実験室への設置まで企図した「電子実験ノートシステム」の構築は、実験中に多数の因子（変数）の動きを固定値として無化しつつ特定の重要な因子（変数）の動きのみにフォーカスする上で、また実験の再現可能性を担保する上でも、必要不可欠である、という話が印象に残った。翻って、歴史学における多様な因子（変数）の取り扱い方と歴史学的実証の「再現可能性」について、改めて自問自答することを強いられた。

文学と科学の間にある歴史学を専攻するにあたって、20～30 代は、あたかも化学式や数式で構築されたような論理明晰で無駄のない客観的な論文

を目指し、19 世紀の実証史家よろしく、科学へと一歩でも近づこうと努力してきた。しかしながら、40 代に入り、再び文学の方向へと戻ろうとしている自分がいる。そうした矢先に、社会実装まで視野に入れたハードサイエンスの真髄をまざまざと見せつけられ、その魅力と真剣さに圧倒された。やはり、母胎に帰ろうと一方に振り切れるのではなく、科学への憧れを真に維持しながら、文学との間で絶妙な均衡を模索していかなければならない。そう思い始めている。

参考文献
[1] 小風尚樹ほか編『欧米圏デジタルヒューマニティーズの基礎知識』、文学通信、2021 年。
[2] 宮川創「ディープラーニングを用いた歴史的手書き文献の自動翻刻:コーパス開発の効率化に向けて」、『KU-ORCAS が開くデジタル化時代の東アジア文化研究:オープン・プラットフォームで浮かび上がる、新たな東アジアの姿』、関西大学アジア・オープン・リサーチセンター、2022 年。

発表者紹介
2017 年、フランス・トゥルーズ第二大学にて博士号（歴史学）取得。2019 年 4 月より大阪市立大学文学部講師。2022 年 4 月より大阪公立大学文学研究科准教授。西洋中世史を専攻し、「13〜14 世紀フランス王国における政府と社会集団の関係」、「13〜14 世紀南フランスの村落共同体における住民自治」などのテーマを、現地で収集した古文書を解読・分析することで探究している。