近代公文書OCR　中京大学 社会科学研究所 山田研究室

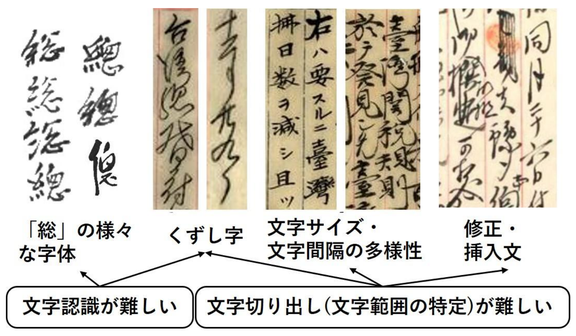
近代公文書とは

- 明治から戦前期までの公文書
- 当時の史実を記録した歴史史料

なぜ近代公文書OCRが必要なのか

- くずし字や古語表現が多く、解読困難な文書
- 行政機関が保管する近代公文書は活用されず死蔵状態化
- 容易に読め、活用しやすくするためOCRが必要
- テキスト化できれば全文検索や翻訳が可能、近代史研究に寄与

【解説】 近代（明治から戦前期まで）の公文書は当時の史実が記録された歴史史料でもあります。しかし、近代公文書は近世の流れをくむ古文書であり、古語的語句・言い回しがあるほか、字体は新旧字体・異体字が混在し、くずし字、略字を使って書かれている場合も多いため、一般の行政職員には解読が容易でなく利活用できず、各行政機関が保管する近代公文書が死蔵状態となりつつあります。また、近代日本史の研究に着手する若手研究者や大学院生にとっても、原史料解読の障壁は高く、いかに容易に読めるようにするかが課題となっています。



台湾総督府文書を題材とした大規模機械学習用データセット

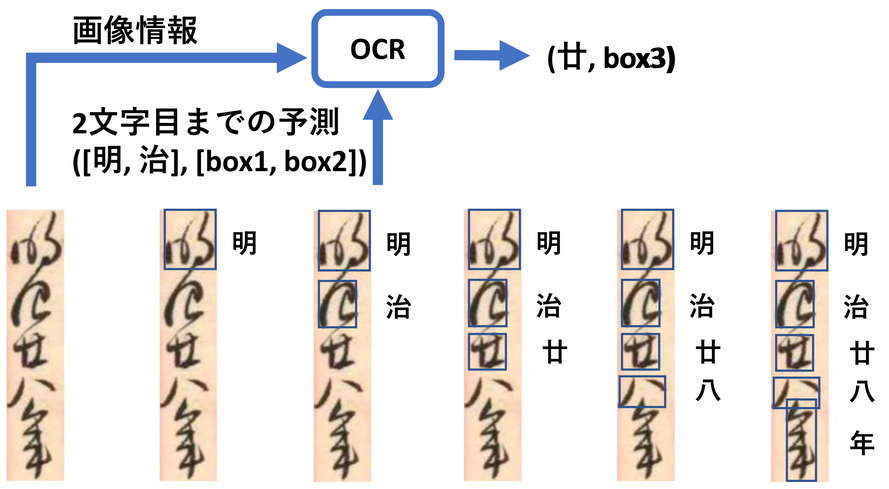
データセット名	言語	統計量			アノテーション			
		原画像数	行数	文字数	行矩形	文字矩形	字種ラベル	読み順
CASIA-AHCDB	中国	11,937		2,276,740			✓	
ICDAR19 HDRC-Chinese DB	中国	1,172	18,752	221,508	✓		✓	✓
TKH dataset	中国	1,000	23471	323,491	✓	✓	✓	✓
MTH dataset	中国	500	17,178	197,886	✓	✓	✓	✓
OCR 学習用データセット（みんなで翻刻）	日本	32,822	572,687	6,791,361	✓		✓	✓
NDLOCR データセット	日本	3,997	238,797	1,348,968	✓		✓	✓
日本古典籍くずし字データセット	日本	6,151		1,086,326		✓	✓	✓
近代公文書データセット	日本	5,003	99,637	1,221,505	✓	✓	✓	✓

- 「台湾総督府文書」からデータ収集
- 近代公文書を対象とした国内最大規模のデータセット
- 文字のバウンディングボックスが付与された高精度データセット

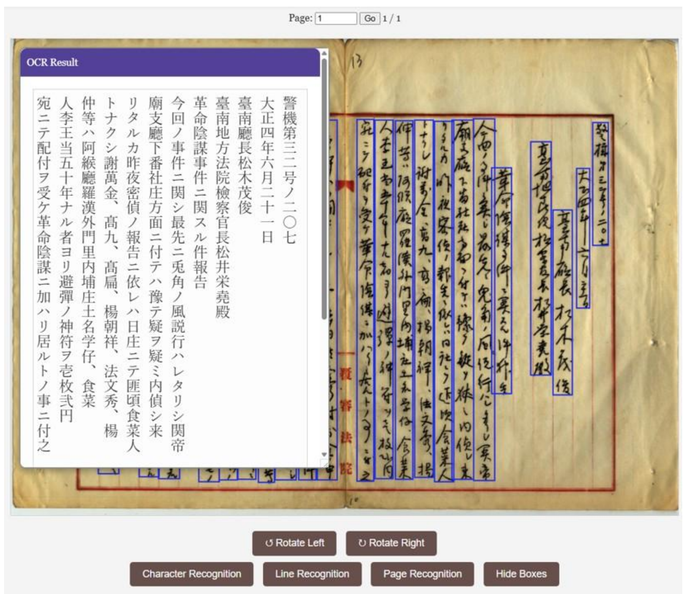
【解説】台湾総督府は日本が台湾を統治していた時代に台湾に設置した行政機関です。台湾総督府文書は明治8年(1895年)から昭和20年(1945年)までのあらゆる種類の公文書（上奏文、法令命令文、内閣文書、各省庁などの関連文書）が原型のままに残された、公文書の雛形的存在で、その量は13,146簿冊（1簿冊約500ページ）にのぼります。台湾総督府文書からサンプリングした5,000ページを日本史の専門家に翻刻してもらい、122万の文字について、個々の文字のバウンディングボックスおよび新旧字体・異体字を区別した字種ラベルを付与した高精度データセットを作成しました。

OCRシステム

- 文書から行ごとに切り出し、行単位で文字認識
- 行内の文字の字種とあわせ、文字のバウンディングボックスも予測することで、認識精度95%を達成



画像情報と既に予測したからn文字目までの字種と位置からn+1文字目の字種と位置を予測



【解説】OCRシステムは、まず文書画像から各行を検出し、それらを整列します。次に、文書画像から切り出した行画像を読み込み、行画像中に書かれている手書き文字を１文字目から順に、どのような文字か（字種）と文字のバウンディングボックスを予測します。予測したn文字目のバウンディングからは、n+1文字目が行画像中のどこから始まるのかがわかり、この情報をシステムに入力してn+1文字目の予測を高精度で行うことができます。この工夫により認識精度が95%まで向上しました。なお、認識精度はIoU0.5としたF1-scoreです。

中京大学　オープンメディアラボ／山田研究室　myamada@sist.chukyo-u.ac.jp
中京大学　社会科学研究所　台湾研究部会

本研究はJSPS科研費JP20H01304、JP23K20105の助成を受けました。また、中京大学戦略的研究事業「デジタル・ヒューマニティーズプロジェクト：日本近代公文書自動解読システムの開発」として中京大学から支援を受け、実施しました。

研究室

研究論文