

市民参加とAI

ー「みんなで翻刻」開発者の立場から

国立歴史民俗博物館 研究部

助教 橋本雄太

みんなで翻刻

仲間と、AIと協力して、
古文書解読に参加しよう！

参加する

※現在新バージョンの試験運用中です。安定稼働するまで
旧バージョンも併せてご利用ください。

 いいね！  ツイート

 B! ブックマーク  LINEで送る



みんなで翻刻

- 概要

- 歴史文献史料の市民参加型翻刻プロジェクト
- 京都大学古地震研究会を中心に開発・運営
- 2017年1月に Web サイトを正式公開
- 地震史料を中心とする災害史料を対象に翻刻

- 新バージョン

- 2019年7月、複数プロジェクトの同時進行に対応した新バージョンを公開
- 翻刻対象を災害史料含む歴史資料一般に拡大し、東寺百合文書などの翻刻を現在進行中

京都大学古地震研究会



- 地震研究や防災研究への応用を目的として古文書を解読
- 地震学，気象学，日本史，地理学，科学史，情報学，図書館司書など多分野の参加



- 2012年の活動開始当初から活発な情報技術活用の試み
- 橋本はD2の2014年から参加

開発の経緯

- 古地震研究会の悩み
 - 歴史地震に関わる研究者は全国の研究機関に点在
 - 遠隔地の研究者も交えて史料を「みんなで翻刻」するための仕組みを作れないか？（代表・中西先生）
- 市民参加型プロジェクトに転換した理由
 1. デジタル人文学分野のクラウドソーシングの流行
 2. 2016年に公開したくずし字学習アプリKuLAの成功
 3. 橋本の博論研究テーマとして

翻刻の対象

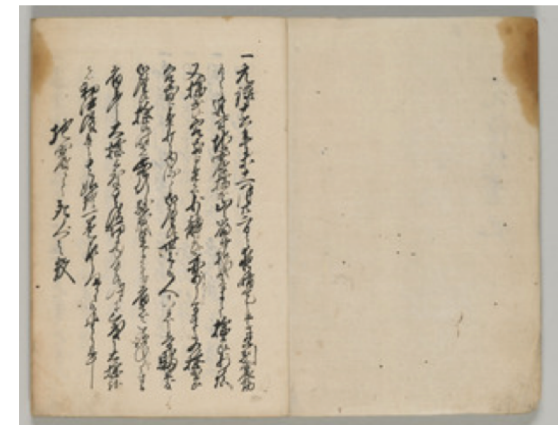
- 東大地震研図書室「石本文庫」
 - 東京大学地震研2代目所長・石本巳四雄（1893-1940）が収集した災害史料コレクション
 - このうち、現在地震研でデジタル公開されている史料114点（画像数3,193枚）の全文翻刻が当初の目標
→5ヶ月弱で完了。地震研所蔵の和古書全点に対象を拡大



『信州浅間山焼』



『地震年代記』



『元禄地震記』

市民参加の実現に向けた課題

1. 規模の問題

- 数百～数千人規模の市民の参加が必要

2. 動機付けの問題

- 善意に頼るだけでは長期的参加は見込めない

3. 能力の問題

- くずし字の正確な翻刻には相当の訓練が必要

『みんなで翻刻』のアプローチ

- くずし字解読の学習サービスとしてデザイン
 1. くずし字解読の学習コンテンツ提供
 2. 参加者間のコラボレーションによる「学び合い」
- くずし字学習の延長として自然に翻刻作業に参加してもらうことが目的
 - ・ 「やりがい搾取」にならないように
 - ・ 参加者が明確なメリットを得る形で参加してもらう

①くずし字学習アプリとの連携

- くずし字学習支援アプリKuLA
 - 阪大文学研究科を中心に開発されたくずし字の学習支援アプリケーション
 - 3,000枚の用例画像と読解教材を収録
 - 2016年2月の公開後13万回DL
- 「みんなで翻刻」との連携
 - KuLA の全コンテンツをサイトに転載
 - くずし字解読の基礎を独習可能に



②参加者間の翻刻文添削

- 翻刻文の保存時に、
翻刻文の添削を他の参加者に
依頼することができる
- 相互添削の目的
 - 参加者の能力向上
 - 翻刻文の品質向上

編集を終了

この内容で保存していいかな？
すべての文字が翻刻できていたら「翻刻完了」にチェックをつけてね。読めない箇所が残っていても大丈夫！
編集コメントを記入しておくと、後の人の助けになるかも！



添削おねがいします

☐ 翻刻完了

☒ タイムラインで共有

☒ 添削希望（他の参加者に翻刻文の添削をお願いします）

キャンセル [保存](#)

翻刻作業のタイムライン共有



みんなで翻刻【地震史料】

はじめに

ホーム

ごあいさつ

ブログ・SNS・ML

地震史料を翻刻する

翻刻ガイドライン

史料一覧

お気に入り・作業履歴

フォーラム

くずし字解読を学ぶ

まなぶ

つながる

ライセンス/クレジット

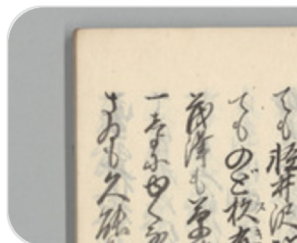
お問い合わせ



田原輝男さんが318文字翻刻しました

1時間前

添削希望



浅間大焼騒動記 中一コマ26

も勤れはて夫より種月然へ押寄大勢ふミ入れ共人壱人も見
へず唯本堂計ぞおわしける然共委...

あっばれ！



ramayakatさんが424文字翻刻しました

2時間前

添削希望



江戸大地震并ニ出火場所細見録一コマ3

一新吉原江戸町京町角町揚屋町伏見丁仲の丁すべてくるのは
こらずゆりつぶれ《？：てみへ出...

あっばれ！

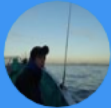


小林じゃさんが679文字翻刻しました

2時間前

添削希望

翻刻文添削時の通知



現在の進捗（日次更新）

総入力文字数	3061194文字 (+12760文字)
翻刻完了画像	4551/4636枚 (98%, +13枚)
翻刻完了史料	263/280点 (+0点)
参加登録者数	3428人

翻刻文字数ランキング

-  Takano Kiyoshi
343211文字
-  田原輝男
322530文字



 Hideo KIHARAさんも【善光寺地震被害...・コマ7を翻刻しました
14時間前

 たかたに やすひこさんがあなたの翻刻に「あっぱれ！」と言っています
14時間前

 レベルが19に上がりました
14時間前

 レベルが19に上がりました
14時間前

 小林じゃさんも伊豆国修善寺温泉...・コマ2を翻刻しました
17時間前

添削結果の表示

地震年代記 ★ 14

閲覧 入力 ノート 編集履歴 凡例

seizai	2017-01-13 08:20:03 - 0文字を翻刻
十一	2017-01-13 06:57:26 - 4文字を翻刻
seizai	2017-01-13 06:50:17 - 626文字を翻刻
大河	2017-01-12 11:09:18 - 130文字を翻刻
十一	2017-01-12 10:49:03 - 60文字を翻刻
大河	2017-01-12 09:48:43 - 12文字を翻刻
十一	2017-01-12 04:19:34 - 93文字を翻刻
steve forrest	2017-01-12 06:09:10 - 40文字を翻刻

十二月廿九日同前《箱…合運図》《箱…明應》二年十月同前《箱…續史
畧》同三年五月七日
同前《箱…合運図》同四年八月十五日、一鎌（かま）くら一大佛地震（だ
いぶつじしん）にて一破損（はそん）。同七年六月
十一日、一天（あめ）の一下（した）一時に大地震。以上《箱…本朝年代
記》同年八月十九日地大に震に此月
しばしば一震（ふる）ひ、一伊勢（いせ）一紀伊（きい）参一河（かは）
の一海（うみ）あふるゝ。《箱…和長郷記》《箱…天正年代記》《箱…後
柏原天皇》
《箱…永正》七年八月七日、一地震（ぢしん）大にとゞまらず、一天王寺
（てんわうじ）一石（いし）の一鳥居（とりゐ）一額（がく）一落（お
ち）、一五畿内（ごきない）
一堂塔（どうたう）一破懷（やれくづ）【懷は壞の誤字でしょうか】（や
れくづ）れ一人（ひと）一多（おほ）く死す。《箱…本朝年代記》同八日
地大に震（ふる）ひ廿八日 遠江（とほく）の
一海（うみ）あふれ、陸地（くがち）一没（おちいり）て海となる。一今
切（いまぎれ）のわたしこれなり。《箱…續史畧》同九年
六月、一大地震（おほぢしん）《箱…合運図》同十一年四月八日同前
《箱…後奈良天皇》《箱…弘治》二年二月
一地（ち）大に震ふ。同三年七月同前《箱…公卿補任》《箱…正親町天
皇》《箱…天正》十三年十一月
十九日《割書…合運図》に廿九日とす 京師（みやこ）地（ち）大（おほ
い）に震（ふる）ひ百日餘（よ）やまず《箱…續史畧》《箱…後陽成天
皇》
《箱…天正》十七年二月駿河（するが）遠江（とほく）ニヶ国
（こく）地（ち）大に震（ふる）ひ民舎（たみのいへ）を懷【壞の誤字で
しょうか】（くづ）す。《箱…武徳大成》

他の参加者から添削を受けた際の修正箇所が一目で確認できる
 緑色＝追記箇所, 赤色＝削除箇所

プロジェクトの進捗（2017年1月～）

参加登録者数 5,346名

総入力文字数 619万文字

翻刻が完了した画像 8,300/8,925枚（92%）

翻刻が完了した史料 498/508点

地震研所蔵のデジタル化済み
和古書資料 全点の翻刻が完了

成果物の品質について

- 中世史研究者に依頼し、翻刻文10万文字を検証：

表 3.2 検証作業を通じて見つかった要修正箇所の数（資料種別・タイプ別）

資料種別	合計文字数	誤読箇所		未読箇所		表記ゆれ箇所		合計	
		実数	比率 (%)	実数	比率 (%)	実数	比率 (%)	実数	比率 (%)
木版本	78,774	400	0.5	285	0.4	395	0.5	1,080	1.3
筆写本	21,901	210	1.0	105	0.5	132	0.6	447	2.0
合計	100,675	610	0.6	390	0.4	527	0.5	1,527	1.5

- 100文字あたり平均1.5文字のエラー
- 学術出版される史料集には及ばないものの、
内容把握や全文検索には十分な品質

「みんなで翻刻」とAI

AI 技術の導入

- 2019年7月にリリースした新バージョンから『みんなで翻刻』にAI技術の導入を進めている
- 市民参加とAIは[互恵的関係を築くことが可能](#)
 1. AIはプロジェクトの参加者を支援し、作業を易化・効率化することができる
 2. 市民参加プロジェクトの成果物は、AIの学習に使用する教師データとして活用できる

くずし字認識AIについて

- 2種類のモデルが利用可能
 1. CODH koguma-net (TensorFlow.js製)
 2. 凸版印刷くずし字認識AI (API経由で利用)
- AIに期待する役割
 - 「補助輪」としてのAI
 - やはり未経験者にくずし字の判読はハードルが高い
 - AIの支援により、初学者が翻刻に参加する敷居を下げる効果を期待

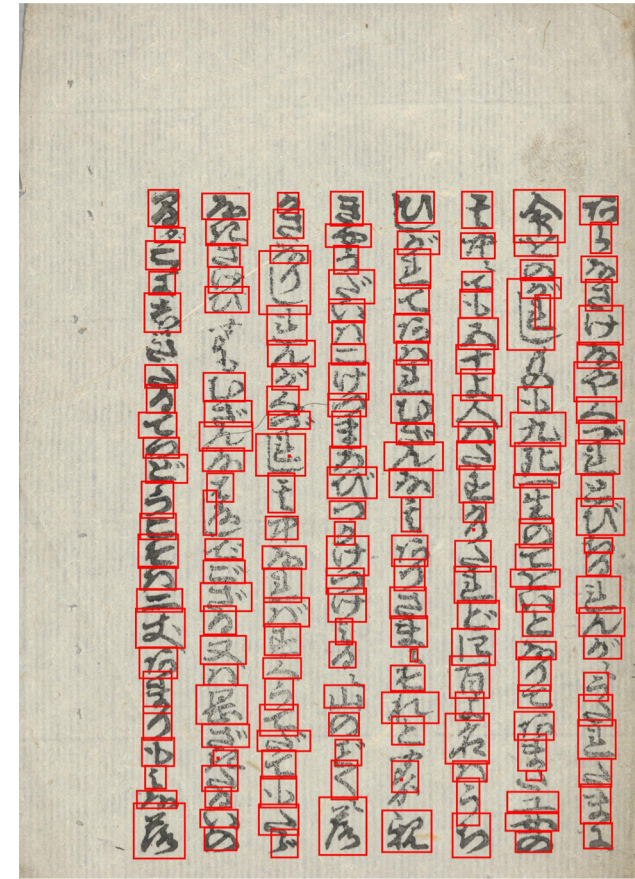
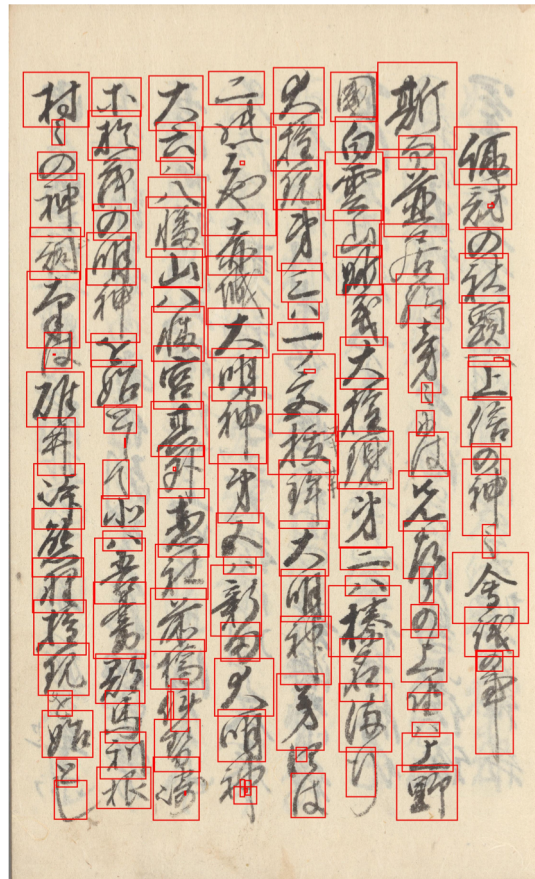
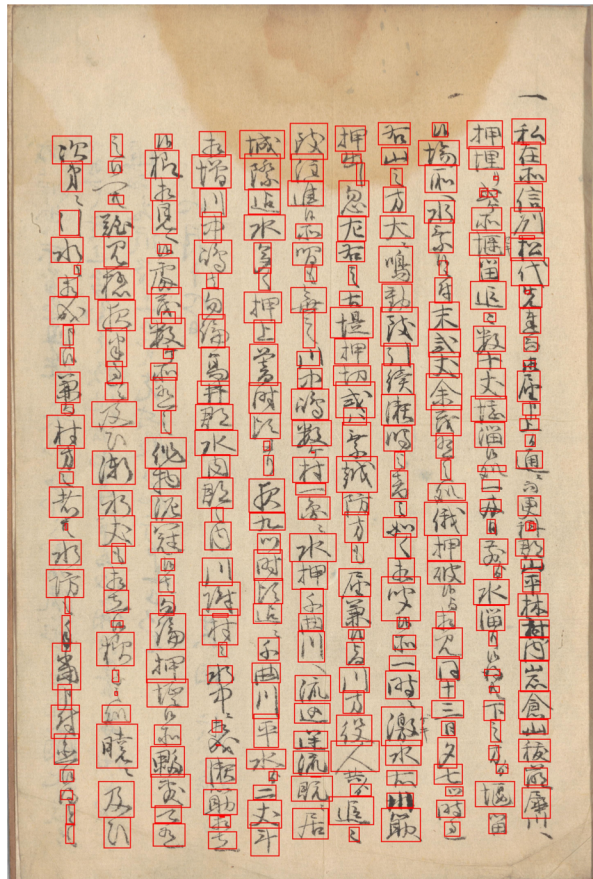
くずし字認識AIについて

- 2種類のモデルが利用可能
 1. CODH koguma-net (TensorFlow.js製)
 2. 凸版印刷くずし字認識AI (API経由で利用)
- AIに期待する役割
 - 「補助輪」としてのAI
 - 学習コンテンツを提供していても、やはり未経験者にくずし字の解読はハードルが高い
 - AIの支援により、初学者が翻刻に参加する敷居を下げる効果を期待

②翻刻文をAIの訓練に使用する

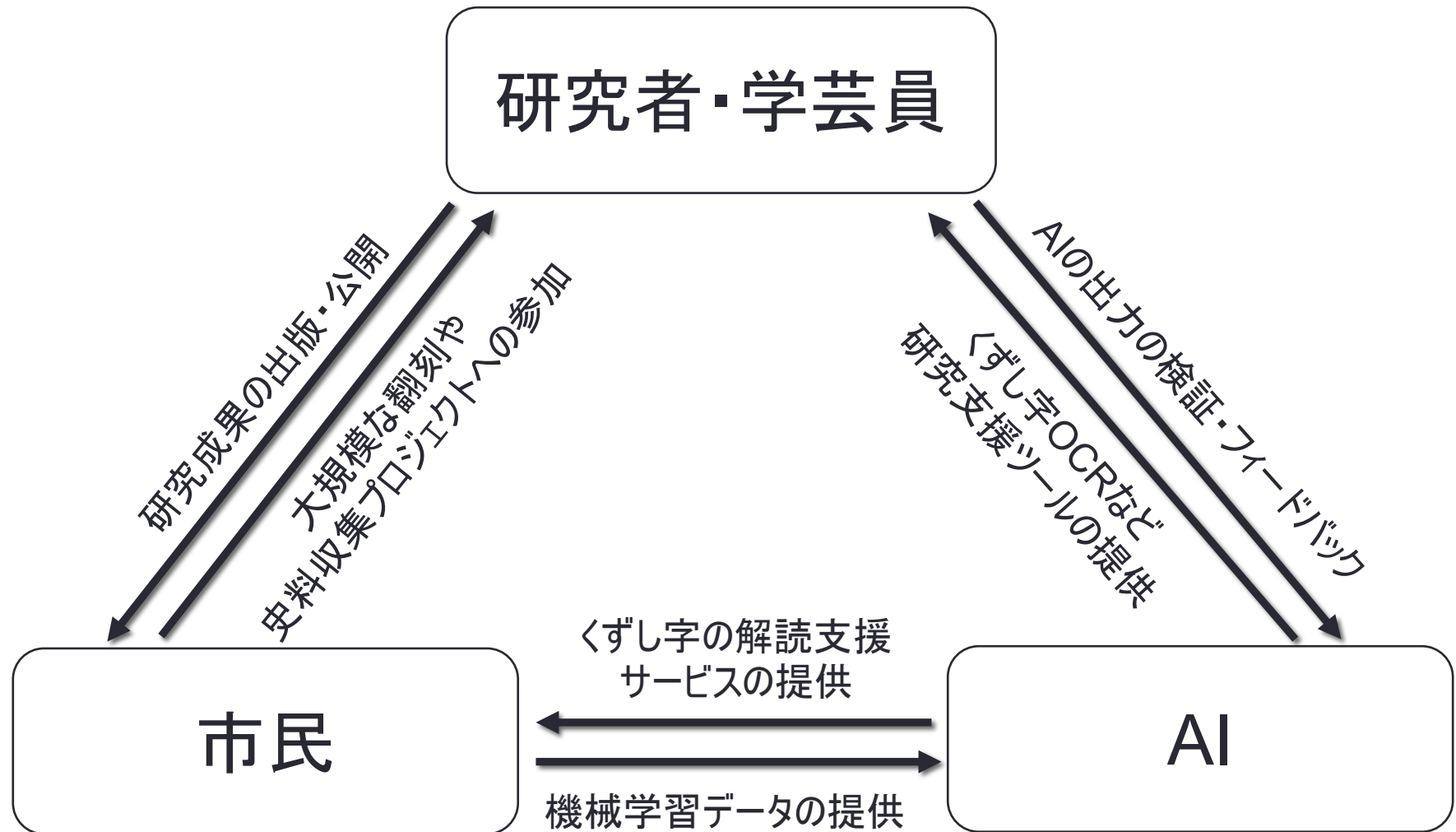
- AIの識別精度は、モデルの訓練に利用するデータセットの分量に大きく左右される
- 市民が翻刻した400万文字（記号類除く）のテキストを機械学習向けに加工できれば、精度向上に大きく寄与する可能性も
- しかし『みんなで翻刻』上のテキストには文字毎の座標情報が含まれない！

KuroNet による文字切り出し



翻刻済みの地震研所蔵史料をAIにより自動で文字切り出したもの
タリン・カラーヌワットさん(CODH)提供

AIと専門家・市民の関係の未来



おわりに

- 専門研究者ではない市民もオンラインで古典籍や古文書を扱う研究に参加・寄与することができる
- AI はこうした人々を代替するのではなく、その能力を強化するために利用できるはず

ご清聴
ありがとうございました