くずし字翻刻のための古文用音声認識システムの検討

研究代表者
文学研究科　戸塚史織
情報理工学研究科　ZHANG Yutao

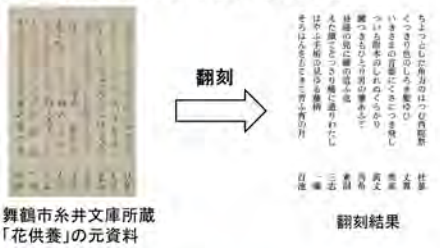
1. 研究背景

□くずし字

- 古典籍、古文書などの**前近代資料に使われてきた文字**
- 日本には数百万の古典籍・古文書があり、近年多くの機関がインターネットを通じてこれを公開し始めているが、くずし字で記されているため現代の多くの人は読めない

➢くずし字翻刻

- くずし字を現在で使用されている文字に変換する



舞鶴市糸井文庫所蔵「花供養」の元資料　翻刻結果

- **負荷が大きい**
- **くずし字解読では近年AIやOCR(光学的文字認識)技術が注目されている**

現状

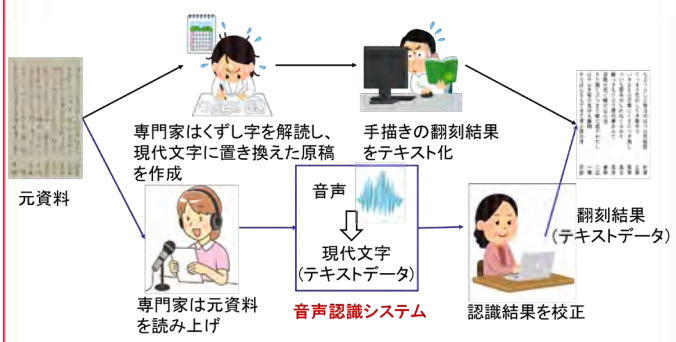
- くずし字解読を得意とする人々の多くは**AIやOCR技術**よりも、使い慣れた**アナログ的な方法**の方がより効率的に翻刻できる
- 翻刻に慣れた人が過去に自分の手元に手書きで残した翻刻の**テキストデータ化**が求められる

2. 古文用音声認識システム

□くずし字翻刻の音声入力システム

- 負荷を減らし、翻刻のテキストデータ化効率化させるため
  - 画像だけからではなく、**音声** からくずし字翻刻する方法を検討
  - 翻刻に**音声認識システムの導入**を検討

➢キーポイント:
高精度な古文認識を実現できる音声認識システムが求められる



□古文用音声認識システムの構築

- 音声認識手法: **DNN-HMM**、**End-to-End**
- DNN-HMM古文用音声認識システム
  - **認識エンジン**: 大語彙連続音声認識エンジンJulius
  - **音響モデル**: ASJ-JNAS コーパスおよび CSJ(日本語話 し言葉コーパス)で学習されたDNN-HMM音響モデルを利用

しかし、一般的な文章の言語データベースから学習された汎用言語モデルと発音辞書は古文に適用された場合、期待するほどの認識精度が実現できない

古文に対して認識精度を改善するため
➢ **古文用音声認識システムの言語モデルと発音辞書を構築**

- 学習データの準備
  - 『日本語歴史コーパス』に含まれる作品
    - ❖江戸時代洒落本(15作品)、文字数:約17万
  - 独自の学習データ
    - ❖役者白虎通 江戸の巻(先頭5000語)
    - ❖**電子辞書の解析結果を元に修正**

聖遊廓	新月花余情	花街寿々女(廓噺余興花街寿々女)	花街鑑(玉菊全伝花街鑑)	異本郭中奇譚
箱まくら(河東方言箱まくら)	色深狭睡夢	風流裸人形	深川新話	原柳巷花語
当世左様候	郭中奇譚	無論里問答	陽台遺編・蜘蛛秘言	月花余情

- 言語モデルの構築
  - 形態素解析結果による
    - ❖学習データの処理:文に分割する、不要部分を削除するなど
    - ❖辞書の構築:語彙へ読みを付与する
  - 処理した学習データからn-gramを推定する
  - バックオフn-gramモデルを構築する

3. 評価実験

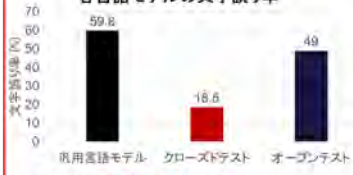
□実験目的

- 現時点で構築した言語モデルの性能を確認

□実験手順

- テストデータ
  - 洒落本の読み上げデータ　内容:「聖遊廓」の一部　長さ:4分くらい
- 音声認識ソフトウェアJuliusを用い、次の三つの言語モデルとその中に含まれた単語で構成された発音辞書でテストデータを認識
  - 汎用言語モデル(『現代日本語書き言葉均衡コーパス』で学習した言語モデル)
  - **クローズドテスト**(テストデータ「聖遊廓」を含んでいる学習データで学習した言語モデル)
    - ❖モデルが学習できたかどうかを確認
  - **オープンテスト**(テストデータ「聖遊廓」以外の学習データで学習した言語モデル)
    - ❖学習データに含まれない古文に対し、認識性能を確認

各言語モデルの文字誤り率



実験結果の詳細

		クローズドテスト	オープンテスト
総文字数	894	894	894
総誤り文字数	533	165	430
文字誤り率	0.60	0.18	0.49
文字認識率	0.40	0.82	0.51

認識結果が**良い例**と**悪い例**

原文	心をやわらぐるもゆかりのつきの一ふしぞかし
汎用言語モデル	心を、和らぐ留萌ゆかりのお付きの人藤井溶かし
クローズドテスト	心をやわらぐるもゆかりのつきの一ふしぞかし
オープンテスト	心を和らぐるもゆかりの月のひとふしぞかし

原文	もし仁さま此中横堀でお見うけ申ましたゆへ
汎用言語モデル	あっぜんさま、この、魔界より。中井を公算に家を三重県申しましたゆえ
クローズドテスト	ます天さまああ此あつて間かゐよりあつて中居を古語にでお見うけ申ましたゆへ
オープンテスト	ます天さまあ此あつて間かゐよりあつて中居をこごりでを身請の申ましたゆへ