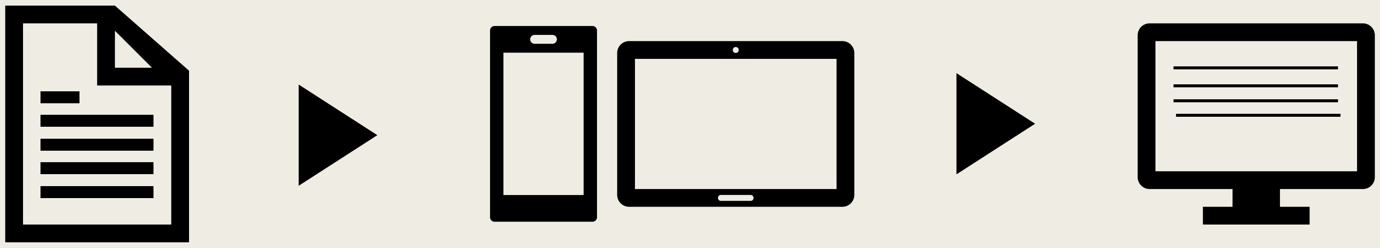
近代書籍テキストの誤字修正

奈良女子大学　高田研究室

研究背景

- 紙メディアのデジタル化→劣化防止，利便性向上
- 現代の主な書籍のデジタル化手法
  - OCR(Optical Character Recognition,光学文字認識)

デジタル化の流れ



印刷した文字
手書き文字

OCRアプリや
専用機器で読み取り

パソコン等で使える
テキストデータに

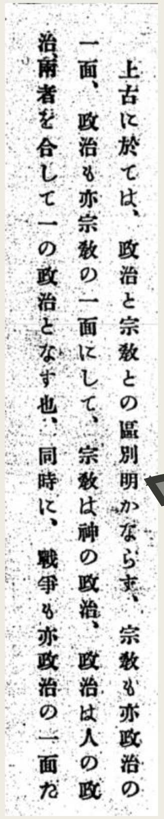
- 近代書籍とは
  - 明治～昭和初期間に日本で出版された書籍
  - 日本における近代化の過程を示す文化的な記録

しかし…

- 近代書籍の特徴
  - 共通規格が未決定
  - 出版者,出版年によりフォントが異なる
  - 活版印刷によるインクのにじみ,かすれが存在

  <u>→現代の書籍を対象としたOCRシステムによる文字認識は困難</u>

■ 近代書籍に特化したOCRも存在するけど…


上古に於ては、政治と宗教との區別明かならず、宗教も亦政治の一面、政治も亦宗教の一面にして、宗教は神の政治、政治は人の政治、兩者を合して一の政治となす也、同時に、戰爭も亦政治の一面た

画像データ

近代書籍に特化した
OCRでテキスト化！

認識ミスが存在
「一面」→「一四」

上古に於ては、政治と宗〓との區別明かならず、宗〓も亦政治の一四、政治も亦宗〓の一面にして、宗〓は、神の政治、政治は人の政治、兩者を合して一の政治となす也、同時に、戰爭も亦政治の一面た…

テキストデータ

- 本研究の目標

  <u>文字認識後の近代書籍に対して文脈を考慮した誤字修正</u>

https://lab.ndl.go.jp/dl/book/798973

～誤字修正のアプローチ～

※現代語の例

①誤文検出

誤りを含む文の特定

- 今日の天汽は晴れだ。
- 明日の予定を考える。
- 友達と食事に行く。

⟹ 今日の天汽は晴れだ。

②誤字検出

誤字を含む単語の特定

- 今日の天汽は晴れだ。

⟹ 今日の天汽は晴れだ。

③誤字修正

文脈から正しい単語を予想

- 今日の？？は晴れだ。

⟹ 今日の天気は晴れだ。

使用技術

- BERT(Bidirectional Encoder Representations from Transformers)
  - 高度な文脈理解が可能な言語モデル
  - 2段階の学習
    - 事前学習

      大量のデータで言葉の使われ方をざっくり学習
    - ファインチューニング

      事前学習済みのモデルに少量の専用データを与えて再学習

ファインチューニングとは…
学習済みモデルを目的に合わせて
調整すること

BERTの事前学習タスク

- MLM（Masked Language Modeling）
  - 入力文中の15%を取得
  - そのうち以下のように単語を置換
    - 80%[MASK]
    - 10%ランダムな単語
    - 10%はそのまま
  - 置換された元の単語を予測するタスク

例）　今日の天気は晴れだ。

置換

今日の[MASK]は晴れだ。

穴埋め問題形式で
[MASK]内の単語を予測

BERTの事前学習タスク

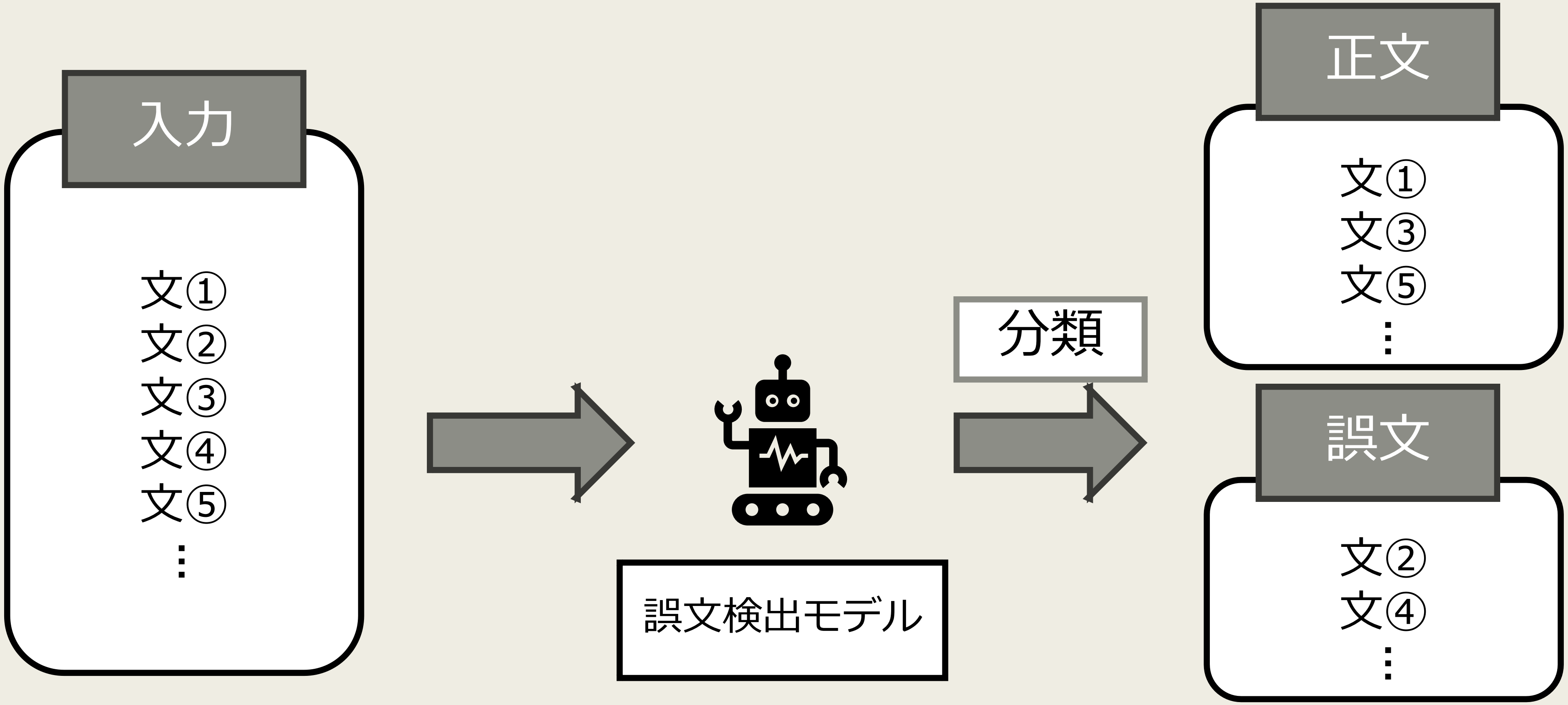
- NSP（Next Sentence Prediction）
  - 入力された2つの文章が連続した文章であるかを学習
  - 50%が自然なつながりのある文章,50%がランダムな文章

2文が連続するかを予測

文A	文B	答え
私は駅に行った。	そこで友人に会った。	○
私は駅に行った。	リンゴは赤い。	×

誤文検出

- 誤文検出にはBERTをファインチューニングしたモデルを使用

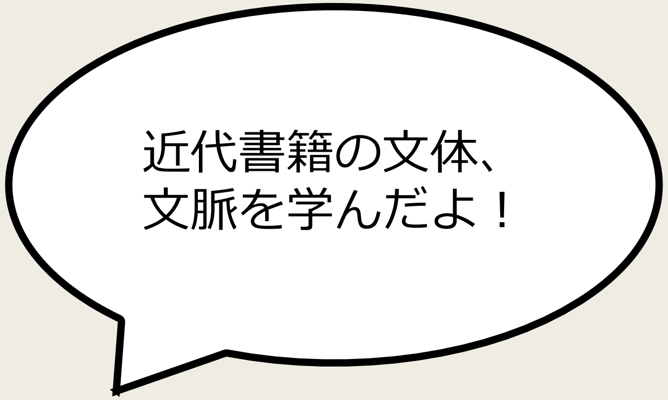


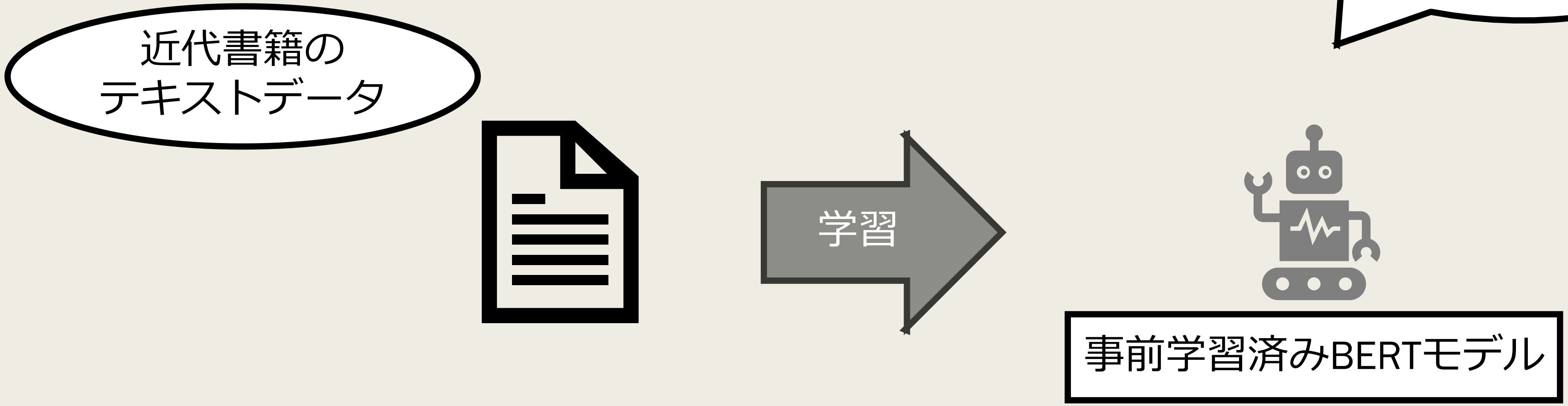
誤文検出のイメージ

- 誤文検出モデル作成手順
  1. BERTの事前学習モデル構築
     - 何のために？
       →モデルを近代書籍にみられる表現に対応させるため
  2. 事前学習モデルをファインチューニング
     - 何のために？
       →モデルが入力文に誤りが含まれるかどうかを判定できる
       ようにするため

誤文検出

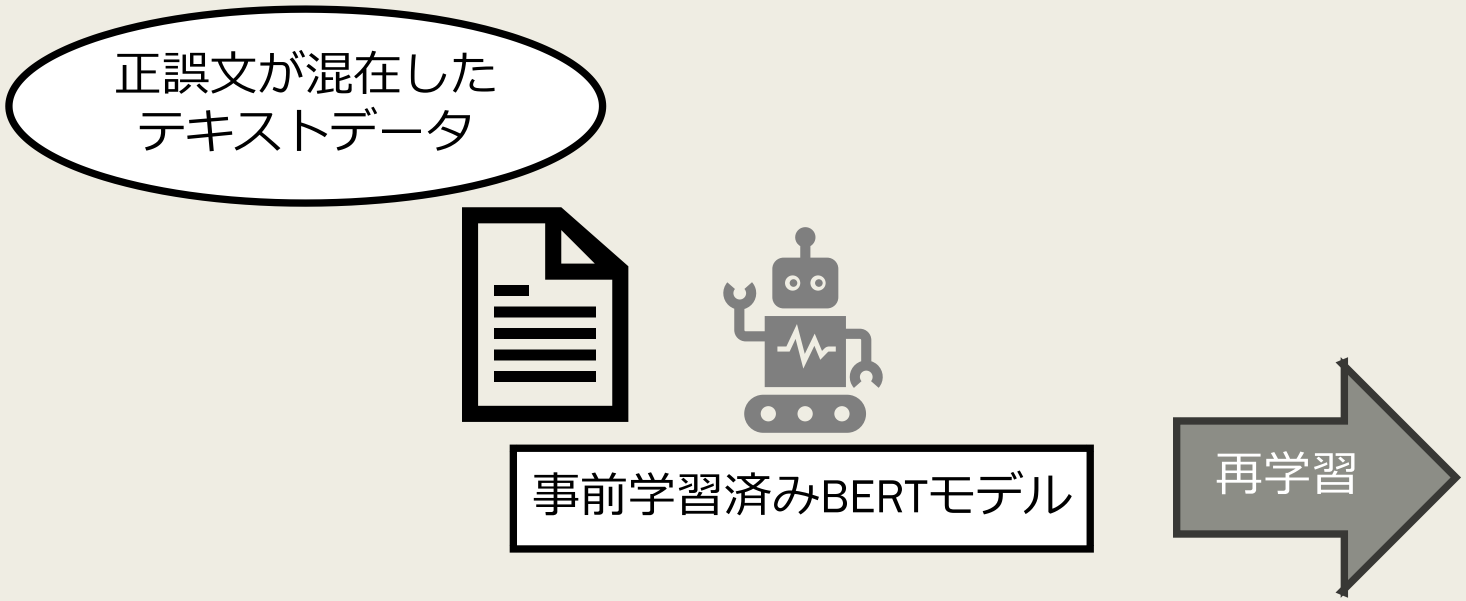
1. BERTの事前学習モデル構築
   - 誤りのない近代書籍のテキストデータで学習
   - 学習タスクはMLMのみを採用
   - トークン化にはSentencePieceを使用

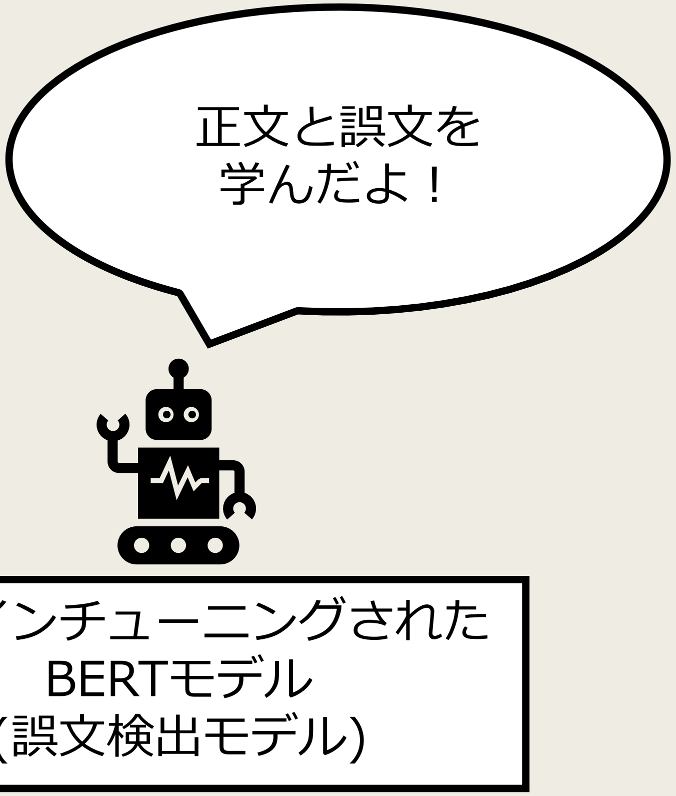




誤文検出

2. 事前学習モデルをファインチューニング
   - 正文と誤文が混在したテキストデータで学習
   - ここで構築されたモデルで誤文検出





■ 誤文の作成方法

① 近代書籍のテキストデータから漢字を収集

['今', '日' ,'勤', …, '題']

② 文を単語分割

['二十五年', '八月', '再び', '樞', '密', '顧問', '官', 'に復', 'した', '。']

3単語以上かつ2文字以上からなる熟語が存在する文を採用

③ ②で選択した文から熟語をランダムに1つ選択, 熟語の2文字目以降を①で収集した別の漢字に置換し，元の文に埋込

['二十五年', '八月', '再び', '樞', '密', '顧問', '官', 'に復', 'した', '。']

顧日

問を日に置換

誤文完成！

二十五年八月再び樞密顧日官に復した。

誤字検出

- 誤字検出には熟語コーパスを使用
  - 近代書籍に登場する熟語を収集した熟語コーパスを作成
  - 誤文検出で検出された誤文に対して実行
  - 誤文中の熟語と熟語コーパスを参照
  - 熟語コーパスに存在しない＝誤字として検出
  - 検出した誤字は[MASK]トークンで置換

入力

二十五年八月再び樞密顧日官に復した。

熟語コーパス

熟語コーパスの中に
「顧日」という単語は存在しないから
「顧日」が誤字！

出力

二十五年八月再び樞密[MASK]官に復した。

誤字検出のイメージ

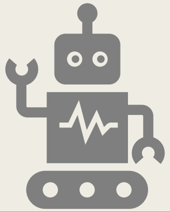
誤字修正

- 誤字修正にはBERTの事前学習モデルとフィルタリングを適応
  1. BERTの事前学習モデル
     - 誤文検出の際に構築したBERTの事前学習モデルを使用
     - [MASK]に入る単語を文脈から推測
  2. フィルタリング
     - 1で推測された単語に対し,誤字と文字列的に近い候補に絞る

入力

二十五年八月再び樞密[MASK]官に復した。

[MASK]に相応しい
単語を上位から予想



事前学習済みBERTモデル

文に相応しいけど
正解とは異なる単語が
予想される場合がある

出力

候補　1位「使」,2位「顧問」,3位「大」,4位「顧」, …

誤字修正のイメージ①

入力

候補　1位「使」,2位「顧問」,3位「大」,4位「顧」, …

元の誤字から最初と最後の漢字を取得
元の誤字「顧日」, 最初の漢字「顧」, 最後の漢字「日」

入力から最初の漢字で始まる or 最後の漢字で終わる単語をフィルタリング

より適切な候補に！

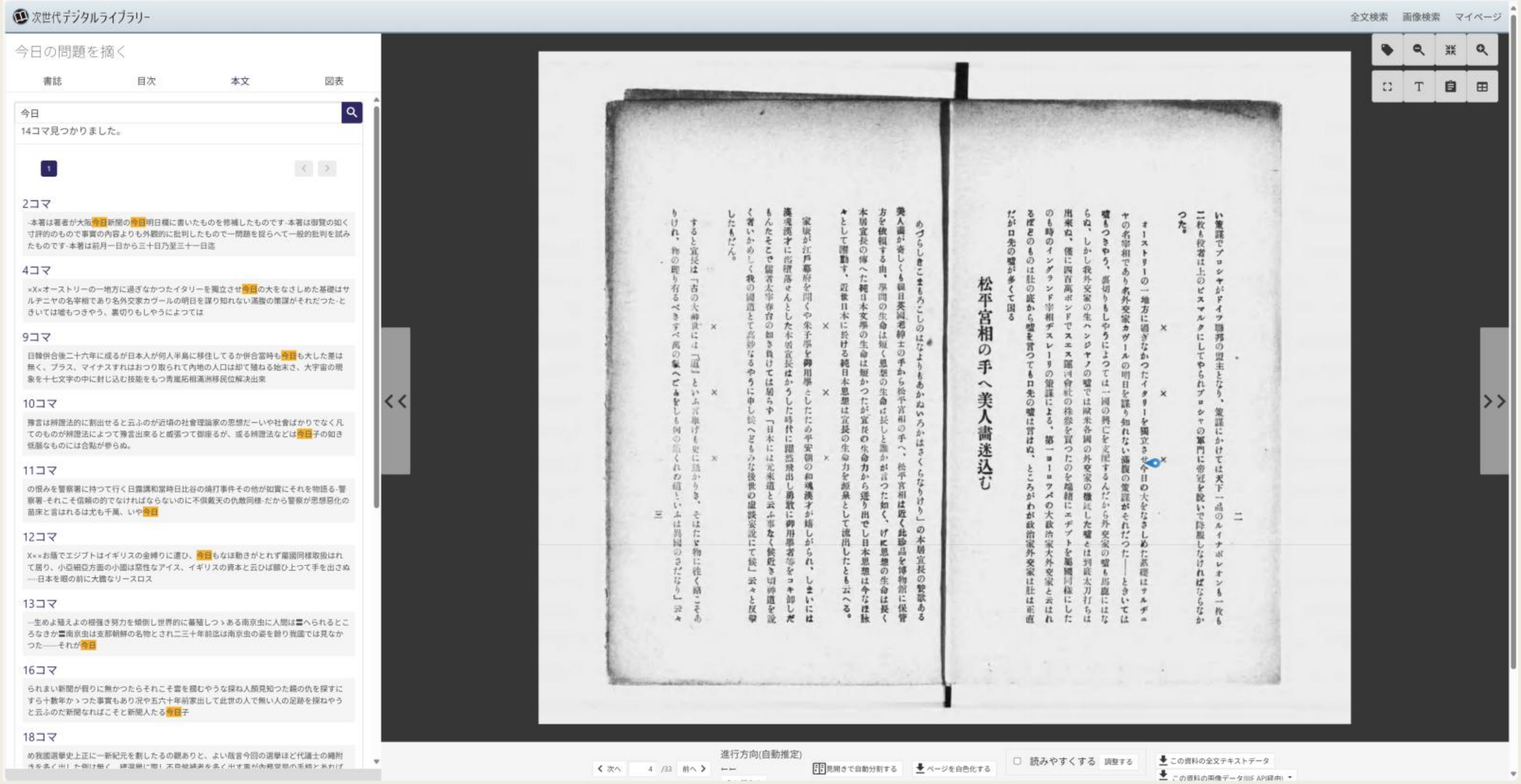
出力

候補　1位「顧問」,2位「顧」,3位「明日」,4位「抗日」, …

誤字修正のイメージ②

使用データ

- 国立国会図書館次世代デジタルライブラリーで取得,正規化を行った近代書籍のテキストデータを使用



次世代デジタルライブラリー：https://lab.ndl.go.jp/dl

学習設定

事前学習

- 学習データ：近代文279,997文
- epoch:130
- batch:32

ファインチューニング

- 学習データ：323,754文 (正文:254,396文, 誤文:69,358文)
- epoch:100

評価手法

- 誤文検出
  - テストデータ3,704文 (正文:2,892文, 誤文:812文)
  - 4つの指標で評価

正解率(Accuracy) $= \frac{TP+TN}{TP+FP+TN+FN}$

適合率(Precision) $= \frac{TP}{TP+FP}$

再現率(Recall) $= \frac{TP}{TP+FN}$

F値 $= \frac{2*Precision*Recall}{Precision+Recall}$

TP：正文であると判断された正文の数
TN：誤文であると判断された誤文の数
FP：誤文であると判断された正文の数
FN：正文であると判断された誤文の数

評価手法

- 誤字検出
  - 誤文検出に成功した誤文に対して実行
  - 誤字箇所を正しく検出できた割合で評価
- 誤字修正
  - 誤字検出に成功した誤文に対して実行
  - 修正候補上位5位内に正解が含まれている割合で評価

実験結果

誤文検出	正解率	適合率	再現率	F値
	94.7%	94.2%	99.4%	96.8%

誤字検出	正解率
	94.3%

誤字修正	正解率
	91.0%

誤文検出，誤字検出，誤字修正ともに9割以上の精度を達成