OCR誤り訂正を用いた歴史新聞データからのコーパス構築

田中 昂志† Chenhui Chu‡ 梶原 智之‡ 中島 悠太‡ 武村 紀子‡ 長原 一‡ 藤川 隆男*
† 大阪大学大学院情報科学研究科
‡ 大阪大学データビリティフロンティア機構
* 大阪大学大学院文学研究科
tanaka.koji@ist.osaka-u.ac.jp {chu, kajiwara, n-yuta, takemura, nagahara}@ids.osaka-u.ac.jp fuji@let.osaka-u.ac.jp

1 はじめに

大規模なテキストコーパスは自然言語処理に不可欠である．既存のコーパスは既に電子化されているテキストから作成されたものがほとんどである．例えば，構文解析のベンチマークである Penn Treebank[*1]は電子化されている Wall Street Journal の新聞記事に対して品詞や構文情報を付与したものである．

一方で，文学をはじめとする様々な分野では，研究対象となる文書の多くが紙などの物理媒体で保存されており，物理媒体をスキャンしたのみで文字起こしなどによってテキスト化されていない．スキャンされた文書画像をテキスト化する技術として Optical Character Recognition（OCR）が存在する．OCR を用いることでテキスト化されていない文書をテキスト化することができるが，OCR は文書の汚れや欠損などが原因でしばしば誤った認識をすることがある [1]．そういった OCR 誤りは文書のコーパスとしての質を低下させる要因となる．したがって，本研究では OCR 誤り訂正を用いたコーパス作成手法の構築を目的とする．

本研究では歴史的新聞データベース Trove[*2]（オーストラリアの主要な日刊紙と地方新聞を網羅）を用いて，19 世紀から 20 世紀の約 120 年間にわたる特定のトピック public meeting に関して記述した記事のコーパスの作成手法を提案する．田中ら [2] の手法では，始めに罫線を検出し，新聞の記事毎に画像をトリミングする．そして，トリミングした記事に対して OCR を適用し，特定のトピックの文字列を含む記事を抽出する．提案手法では，得られた OCR 結果に対して OCR 誤り訂正を適用し，訂正したテキストに対して記事抽出を行う．正解データを人手で作成し，評価を行った結果，特定の対象となる記事の始めの文と終わりの文の言語的特徴を用いた手法と比較して，過不足なく抽出できた記事の割合が 15.9% 向上した．また提案手法において，OCR 誤り訂正を適用していない場合と比較して，F 値において 2.5% 向上し，OCR 誤り訂正の有効性を示した．

2 提案手法

提案手法の全体図を図 1 に示す．提案手法では，まず新聞画像内の罫線を検出し，トリミングすることで記事画像を抽出する．次に，抽出した全ての記事画像に対して OCR を適用してテキストを抽出し，得られた OCR テキストに対して OCR 誤り訂正を適用する．そして，検索文字列が含まれているか否かでフィルタリングを施すことで対象となる記事を抽出する．

2.1 トリミング

新聞画像内の罫線の検出及びトリミングには OpenCV を使用する．まず大津の 2 値化法 [3] を用いて新聞画像を 2 値化し，輪郭追跡処理アルゴリズム [4] により 2 値化された画像の輪郭を抽出する．そして，閾値以上の高さと閾値以下の幅を持つ領域をその新聞画像におけるカラムと判定し，閾値以上の幅と閾値以下の高さを持つ領域をその新聞画像における記事区切りと判定する．閾値は人手でチューニングを行い，トリミング操作で得られた画像を記事画像とする．

2.2 OCR

OCR は一般的に文字の区切りの認識，大きさの正規化，特徴抽出，分類という手順で行われる．Google Drive[*3]の OCR と Tesseract の OCR を試した結果，の方が精度がよかったため，本研究では記事画像から

[*1] https://catalog.ldc.upenn.edu/LDC99T42
[*2] https://trove.nla.gov.au
[*3] https://www.google.com/intl/ja_ALL/drive/

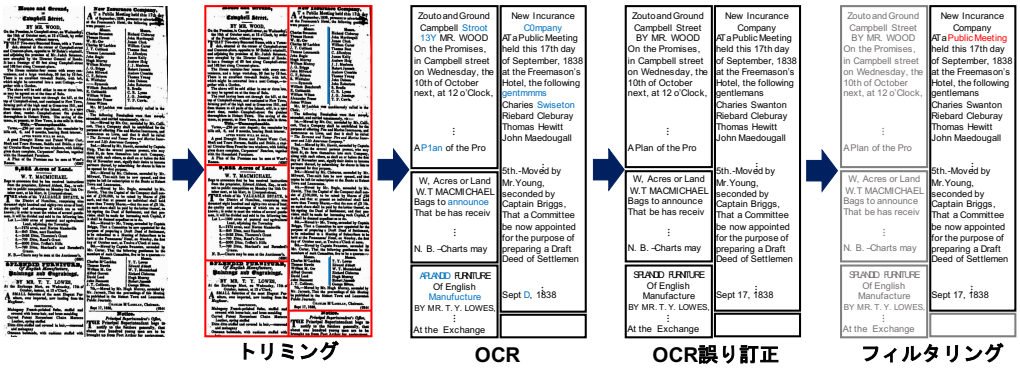

図 1: 提案手法の全体図

	記事数	行数	WER(%)	CER(%)
訓練用	577	11,575	26.50	9.82
開発用	71	1,227	25.49	9.66
評価用	71	1,389	26.57	9.68

表 1: OCR 誤り訂正の教師データの統計

テキストを抽出するために，Google Drive の OCR 機能を使用する．

2.3 OCR 誤り訂正

提案手法では，OCR 誤り訂正手法をオーストラリアの歴史新聞ドメインに特化させるために，OCR 誤り訂正の教師データを作成し，作成した教師データを用いて OCR 誤り訂正手法を学習させる．

2.3.1 教師データ作成

1838 年から 1954 年に刊行された新聞を各年 5 つ Trove からサンプルし，サンプルした新聞内から「public meeting」を含む記事のテキストを抽出する．テキストは Trove の OCR 結果を使用する．オーストラリア史の専門家指導の下，5 人のアノテータによって新聞画像を見ながら OCR 誤りをしているテキストを訂正する．したがって，OCR 誤り訂正前のテキストと人手による OCR 誤り訂正後のテキストのペアが得られるので，このテキストのペアを教師データとする．表 1 に作成した教師データの統計を示す．ここで，Word Error Rate（WER）は正解文と比較して OCR 文が誤っている単語の割合を示し，以下のように計算される．

$$\mathrm{WER} = (S + I + D)/N \quad (1)$$

S は置換された単語の数，I は挿入された単語の数，D は削除された単語の数，N はテキストに含まれる単語の数を示す．Character Error Rate（CER）は WER を単語単位ではなく文字単位で計算する．

2.3.2 OCR 誤り訂正手法

OCR 誤り訂正手法として，統計的機械翻訳（SMT），ニューラル機械翻訳（NMT），半教師あり学習手法を用いて実験した結果，SMT が最も高い精度を示したため，提案手法では SMT を用いる．SMT ツールとして Moses[*4]を使用し，作成した教師データを用いて文字単位で学習させる．評価用データで評価した結果，WER にて 3.14%，CER にて 0.61% となった．

2.4 フィルタリング

記事画像の OCR 結果に検索文字列が存在するか否かでフィルタリングを行い，対象となる記事を抽出する．OCR による文字認識の誤りを許容するため，文字単位での類似度を条件とする．類似度の計算には Python の difflib モジュールの SequenceMatcher[*5]を使用する．SequenceMatcher は以下のように文字列間の類似度を計算する．

$$Similarity = \frac{2 \times M}{T} \quad (2)$$

M は一致する文字数，T は比較する文字列の合計文字数を示す．検索文字列の単語数に応じた単語 N-gram を記事内のテキストから得る．得られた N-gram と検索文字列との類似度を計算し，類似度の最大値が閾値以上の記事を対象記事とする．閾値は開発用データでの評価で最も F 値が高い閾値を用いる．

3 評価実験

3.1 使用データ

本実験では，1838 年から 1954 年に Trove で刊行された「public meeting」を含む新聞 307 件を開発用 149

[*4] http://www.statmt.org/moses/
[*5] https://docs.python.jp/3/library/difflib.html

件，評価用 158 件にランダムに分割して使用する．対象記事の正解データは人手で抽出したものを用いる．

3.2 比較手法

Baseline Trove の OCR テキストに対して OCR 誤り訂正を適用したテキストの特徴量から記事の特定を行う手法を用いる．Baseline では，対象記事の初めの文と終わりの文の特徴をそれぞれ検出し，記事の特定を行う．Baseline で用いる文の特徴は以下の通りである．

始めの文 「public meeting」を含む文から前 2 文を取得．

終わりの文 Stanford parser[*6]を用いて固有表現抽出を行う．固有表現タグ（LOCATION, DATE, PERSON）を含む文かつ次の文が固有表現タグを含まない文を取得．

Baseline+Proposed 始めの文に対応する終わりの文が検出できず，Baseline で記事抽出に失敗した新聞に対して Proposed を用いる．

w/o Correct コーパス構築手法における OCR 誤り訂正の有効性を検証するために，Baseline，Proposed，Baseline+Proposed において，OCR 誤り訂正を適用しない手法を用いる．

3.3 パラメータチューニング

2.1 節で言及したトリミングにおける罫線判定及び小カラム判定の閾値は，1838 年の 1 カ月の新聞データに対して実験的にチューニングを行う．また 2.4 節で言及したフィルタリングにおける閾値は，開発用データを用いて閾値を 0 から 1 まで 0.05 区切りで変化させ実験した結果，F 値が最大となった 0.8 を用いる．

3.4 評価手法

実験では，対象記事が抽出できているかを評価する記事レベルの評価と，記事の抽出の精度を評価する行レベルの評価を行う．また，正解となる記事は OCR 誤り訂正を適用した記事を使用する．それぞれの評価の方法は以下の通りである．

記事レベルの評価手法

抽出された記事内の「public meeting」を含む文と対象記事内の「public meeting」を含む文の類似度を計算し，類似度が閾値以上の場合抽出成功，閾値未満の場合抽出失敗とする．閾値は文として類似している実験的な値として 0.6 を用いる．Baseline，Proposal それぞれに対して評価し，Precision, Recall, F 値を計算する．

行レベルの評価手法

抽出された記事と対象記事の始めの行と終わりの行それぞれに対して調査を行う．抽出された記事が対象記事と比較して余剰・不足している行の割合を計算する．

[*6] https://nlp.stanford.edu/software/lex-parser.shtml

手法	Prec.	Rec.	F 値
Baseline w/o Correct	67.6	60.8	64.0
Baseline	**71.1**	63.9	67.3
Proposed w/o Correct	59.4	51.9	55.4
Proposed	61.9	54.4	57.9
Baseline+Proposed w/o Correct	53.1	91.1	67.1
Baseline+Proposed	54.2	**93.7**	**68.7**

表 2: 記事レベルの評価

3.5 結果

記事レベルの評価

表 2 に記事レベルの評価結果を示す．Baseline+Proposal が最も高い F 値を示した．よって，Baseline で抽出に失敗した public meeting の記事を Proposal が正しく抽出できていることがわかる．また，Proposed より Baseline の方が高い F 値を示した．Baseline は記事の始めの行を取得するとき「public meeting」が含まれている文という特徴を使用しているため，記事レベルの評価の基準となる文を必ず含むように抽出していることが原因であると考えられる．ここで，Baseline が抽出に失敗している記事が存在する理由として，1 つの新聞画像に複数の public meeting の記事が存在し，public meeting が OCR 誤りしている記事が含まれているためである．また，いずれの手法においても OCR 誤り訂正を適用しない手法より OCR 誤り訂正を適用した手法の方が高い精度を示した．したがって，Baseline 及び Proposed において OCR 誤り訂正はコーパス構築に有効であることがわかる．

行レベルの評価

図 2 に始めの行の評価結果，図 3 に終わりの行の評価結果を示す．横軸が余剰，不足している行の割合，縦軸が記事の数を表す．図 2, 3 より，記事

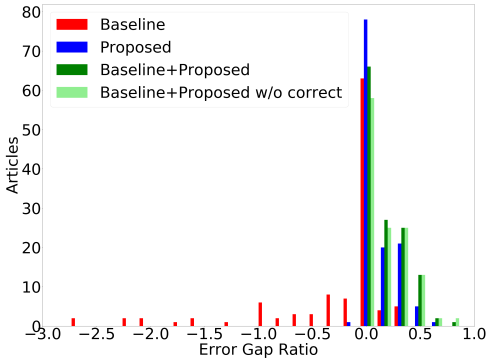

図 2: 行レベル評価（始めの行）

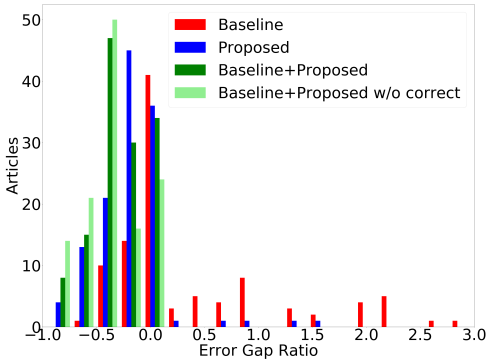

図 3: 行レベル評価（終わりの行）

の始め・終わりいずれの行においても，Baseline と比較して Proposed の方が過不足なく抽出できている記事が多いことが分かる．また，始めの行と終わりの行共に過不足なく抽出できている記事の数は，Baseline では 13 件（8.2%），Proposed では 38 件（24.1%），Baseline+Proposed では 21 件（13.3%）であった．よって，画像の特徴を用いて記事の区切りを検出する手法が，特定の記事を抽出する場合に有効であることが示せた．また，図 2，図 3 より，記事の始め・終わりいずれの行においても Baseline+Proposal w/o correct より Baseline+Proposal の方が過不足なく抽出できている記事が多いことがわかる．したがって，行単位の精度においても OCR 誤り訂正は有効であることがわかる．

3.6 考察

表 3 の上の例は Proposed において OCR 誤り訂正を適用した結果，抽出に成功した例である．誤り訂正前では「publie meetide」と OCR 誤りしており，抽出に失敗していたが，OCR 誤り訂正を適用することで，「publie」を「public」と訂正できたことで抽出に成功している．このような例は 6 件（3.8%）存在した．一方

誤り訂正前	誤り訂正後
usual Publie Meetide of # PUILIO MEETING of	usual Public Meetide of #PUILIOMEETING of

表 3: 誤り訂正により抽出成功及び失敗した例

で，表 3 の下の例は Proposed において OCR 誤り訂正を適用した結果，抽出に失敗した例である．誤り訂正前では「PUBLIO MEETING」と OCR 誤りをしているが「public meeting」との類似度は高い．しかし，OCR 誤り訂正を適用することで，「#PUBLIOMEETING」と訂正したことにより「public meeting」との類似度が低くなった結果，抽出に失敗している．このような例は 2 件（1.3%）存在した．また，Baseline+Proposed で抽出失敗した記事は，Baseline の特徴に合致せず，トリミングでカラムを跨いで切り取っていた場合であった．このような例は 10 件（6.3%）存在した．したがって，Proposed のトリミング精度を高めることでさらに抽出の精度を向上できると考えられる．

4 まとめ

本研究では，歴史的新聞から特定のトピックの記事を抽出し，コーパスを作成する手法を提案した．Trove から取得した 1838 年から 1954 年の新聞データを用いて評価実験を行った結果，評価用データの 93.7% の記事から対象となる記事を抽出し，さらに 13.3% の記事が過不足なく抽出可能であることを示した．今後の課題として，抽出した記事を解析するために，public meeting の開催日時や場所などの情報抽出に取り組む予定である．

本研究は，科研費#19H01330 の助成を受けたものである．

参考文献

[1] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. Impact of ocr errors on the use of digital libraries: Towards a better access to information. In *Proceedings of the JCDL '17*, pages 249–252, Jun. 2017.

[2] 田中 昂志, Chenhui Chu, 中島 悠太, 武村 紀子, 長原 一, 藤川 隆男. 歴史新聞データからのコーパス構築. **言語処理学会第 *25* 回年次大会**, Mar. 2019.

[3] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, Jan 1979.

[4] Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.