証憑を用いた日本語 OCR 誤り訂正ベンチマークの構築

藤武 将人
ファーストアカウンティング株式会社 FA Research
fujitake@fastaccounting.co.jp

概要

この論文では OCR(Optical Character Recognition) システムにおける日本語証憑の認識結果の誤り訂正手法検討のため, ベンチマークとベースラインの構築及びその有効性の評価を行う. 請求書などの証憑において, 適切な処理やその自動化に向けて記載されている文字を正しく認識し, また誤りがある場合には補正を行うことは重要である. そのため, 本研究では日本語請求書における社名などの項目を対象に既存の OCR モデルによる文字認識精度を測定し, 誤り訂正ベンチマークを構築した. そして, 言語モデルを用いた誤り訂正手法を提案し, これらの誤りを効果的に訂正できるかを検証した. 提案した誤り訂正アルゴリズムは, 全体的な認識精度を著しく向上させることができることを示した.

1 はじめに

ビジネス・オートメーションの場面では, 請求書のような企業文書画像からテキストを正確に抽出するシステムが必要とされることがある. デジタル文書が普及したとはいえ, 紙ベースの証憑はまだ存在するため, スキャンした証憑からテキストを読み取る OCR（Optical Character Recognition）技術 [1, 2] は欠かせない. その必要性から様々な OCR サービス [3, 4] や高精度化に向けた研究 [5, 6] が近年進められている. しかし, OCR は様々なフォントや画像の状態に影響され, 正確な読み取りができないことがある. また, 日本の会計実務では, 処理の信頼性を高めるために, 社名などの上に印鑑を押す習慣がある. その結果, 図 1 のような文字画像が作られることがあり, 印影の影響で認識精度が著しく低下する傾向がある. そのため, 正確な情報を抽出するためには, 読み取り結果の誤りを補正する必要がある.

日本語 OCR の誤り訂正は, 書籍や古文書で研究されてきた. これまでの研究は, 誤り箇所の検出と誤り候補の訂正の 2 つの要素から構成されている. 検出では, トリグラムを用いて各文字が誤りかどうかを判定し, 言語モデルにおける確率値が閾値より小さい場合, その文字を誤り候補とする. そして, 誤り候補を訂正する際には, 誤り訂正候補の各文字位置に対して出現確率の高い文字列を置換する [7]. 修正においては, その OCR のテキスト全体のグローバル情報を用いて修正を行うなどの方法が提案されている [8]. 先行研究は, 主に書籍や古文書の書き写しを対象としており, 企業文書や, そのような補正のためのオープンな利用可能なベンチマークの構築には至っていない. また, 先行研究は閾値の設定など複雑で人手に依存している.

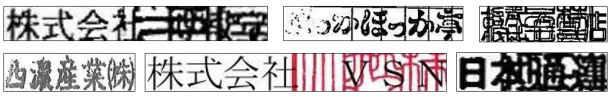

図 1 証憑画像における文字認識対象のサンプル例. 証憑に印鑑を押す慣習により, 日本の証憑は OCR が難しい画像になる傾向がある.

これらの問題に対処するため, 我々は日本の請求書に記載された取引先名に焦点を当て, 日本語の OCR 誤り訂正ベンチマークを構築した. 証憑には会社名, 金額, 日付など様々な情報が含まれており, それぞれに OCR 誤りの可能性がある. しかし, 金額や日付などの情報は 1 つの情報として言語的な意味を持たず, 効果的な誤り訂正は困難と考えられる. そこで, 本研究では取引先名である会社名に着目し, ベンチマークを構築した. そして, そのベンチマークを用いて, 言語モデルとルールベース手法を用いたベースラインを構築し, 評価を行った. 我々の貢献は以下の通りである：

- 日本の証憑に基づく OCR 誤り訂正ベンチマークを提案する. これは 2 つの OCR サービスに基づく実際の誤りであり, 異なる特徴を提供し, 今後の研究の基礎となる.
- 言語モデルとルールベースの解析手法を用いたベンチマークのベースラインを提供した.

表 1　各種モデル・サービスによる OCR 認識精度

Method	Acc w/o preprocess	Acc w/ preprocess
Japanese OCR [12]	24.6	26.6
Google Vision [3]	42.6	72.0
Fast Accounting Robota [4]	97.0	99.1

2　関連研究

2.1　日本語 OCR 誤り訂正

日本語 OCR 誤り訂正タスクは書籍などを対象として研究されており [7], また, グローバル情報を用いて修正候補を選択する改善方法が提案されている [8]. また, 言語モデル BERT を用いた OCR 誤植検出手法も提案されている [9]. 本研究は, 従来の研究とは異なり, OCR 誤り訂正タスクの誤り検出と訂正を単一の言語モデルで直接行うアプローチを提案し, またベンチマークを作成することで, 今後の研究に役立てる.

2.2　事前学習済みモデル: T5

近年の研究では, 大量のラベルなしデータを用いて訓練され, 多くのタスクで良好な性能を発揮する事前訓練モデルに注目が集まっている. これらのモデルの一つである T5 [10] は, 自然言語処理の多様なタスクを系列から系列への変換プロセスとして捉え, Transformer [11] を活用することで性能を向上させている. 本研究では, OCR 誤り訂正を系列から系列への変換と定義し, T5 を用いてこのタスクをモデル化する.

3　ベンチマーク構築

3.1　OCR 評価

日本語証憑における印影の影響を含む OCR 誤り訂正データセット構築を行なった. 我々は許諾をいただいた日本語請求書を用いて, 図 1 に示すように, 社名項目部分について, 文字領域を切り取り画像セットを準備した. ランダムにサンプリングし, 11,000 枚の証憑の取引先画像を用意した. そして, その画像についてアノテーションデータと実際にその画像を OCR モデルが文字認識した結果のペアを作成した. 具体的には, アノテーションは文字認識画像のアノテーションを専門的に行なっている人間により全ての画像に対して Ground Truth(GT) を付与した. OCR については, 現在のテキスト認識のパフォーマンスを評価するために, 3 つのモデルとサービスを使用した. 1 つ目は Japanese OCR [12] で, オープンに公開されている Transformer ベースの日本語テキスト認識モデルである. 2 つ目は Google が提供する Vision API [3] で, 商用汎用テキスト認識である. 3 つ目はファーストアカウンティング株式会社が提供する会計特化の OCR である Robota API [4] である. 各モデルの文字認識精度は, 単語レベルの精度で評価した. つまり, GT と OCR 結果のすべての文字が一致した場合, True となり, そうでない場合は False となる. アルファベット, 英数字, スペースは半角文字として扱った. GT と各モデルのテストセットの文字認識結果例を表 1 に示す. 空白文字, 改行文字, 文字コードの認識はモデルやサービスによって異なる. 空白, 改行などは重要な意味を持つ場合もあるが, OCR 誤り訂正を考慮するとテキスト情報のみから補完を行うことは困難であるため, それらを削除する等の標準化処理を行なった精度を別途算出した. Robota API, Vision API, Japanese OCR の順に, 統一処理を加えた後の OCR 精度は, それぞれ 99.1%, 72.0%, 26.6%となった.

また, 表 2 にランダムサンプルの GT とその OCR 結果を示す. Japanese OCR は日本語の文章に似た認識結果を返す傾向があり, ノイズなどの影響で正しく認識できないことが確認できる. また, Vision API では, 読める文字は正しく認識されるが, 一部の文字が欠落している場合がある. Robota API では欠落文字が少ないことが確認できる.

3.2　OCR 誤り訂正ベンチマークの構築

これまでの結果をもとに, OCR 誤り訂正ベンチマークを構築した. 本研究における OCR 誤り訂正タスクは, 誤りを含む可能性のあるテキスト列を入力とし, 補正されたものを出力として出力することである. 入力として誤りがなければ, 正しいテキストがそのまま出力される. 以上のタスク設定に基づき, テキスト認識結果と GT を用いてベンチマークを構築した. 各モデルの認識結果が入力となり, 出力が GT として推定される. また, 人名のサンプルからは個人情報を除去した.

ここで, Japanese OCR は精度が低く, 定性的にも出力結果から正しい文字を推定することは困難であると考えられたため, ベンチマークデータには Vision API と Robota API の結果をそれぞれのセットとして用いた. 従って, ベンチマークは 2 つのセットで構成

表 2　OCR 認識結果の例

Ground Truth	Japanese OCR	Vision API	Robota API
向島運送株式会社	河島選選が出てきます.	向島運送株会社	向島運送株式会社
株式会社 DAISHIZEN	株式会社 DAISHIZEN	#DAISHIZEN	株式会社 DAISHIZEN
日本ロジテム株式会社	日本の日が流れましたが,	日本株式会社	日本ロジテム株式会社
株式会社横浜ファーマシー	株式会社機浜ファーナー	株式会社横浜ファーマジ	株式会社横浜ファーマント

表 3　訓練・評価・テストセットのデータサンプル数

Set	Vision True	Vision False	Robota True	Robota False
Train	6,817	2,640	9,371	86
val	353	145	490	8
Test	717	279	983	13

表 4　OCR 誤り訂正精度

Method	Vision set	Robota set	Average
Rule-based	76.2	98.1	87.2
$T5_{Megagon}$	12.7	5.9	9.3
$T5_{Retrieva}$	90.1	99.5	94.8
w/o correction	72.0	98.7	85.4

される. 評価は OCR 認識精度評価と同じ, ワード単位精度を用い, それぞれのセットにおける評価及びそれらの平均値を計測する. 訓練, 評価, テストセットとして画像に対して, ランダムに分割を行なった. 表 3 に, 各セットにおける訓練, 評価及びテストの True と False のサンプル数を示す. True は GT と OCR の結果が一致したことを示し, False は誤りを含むサンプルである. Robota セット, Vision セットでは共通の画像を有しているが, それぞれの認識結果精度が異なるため, True と False の割合が異なる. このようにして, 割合や誤り傾向の異なる 2 つのセットを用いて, 日本語 OCR 誤り訂正ベンチマークを構築した.

4　ベースライン構築

4.1　T5 によるアプローチ

OCR 誤り訂正に対して, 言語モデルがどの程度の誤り訂正ができるかについてその性能を評価する. その誤り訂正における評価に適したモデルとして T5 を用いて実験を行った. T5 は Transformer をベースとしたエンコーダーデコーダー型の言語モデルであり, その構造から, 系列から系列へと変換が可能である. そのため, 本研究のタスク設定である, 誤りが混入しているかもしれない文字列を入力として, 訂正後の出力が可能となる. 我々は, 2 つの日本語の事前学習済みモデルを使用した. そして, その事前学習済みモデルに対してベンチマークで微調整訓練を行なった.

一つは Megagon Labs が公開している学習済みの日本語 T5 であり, 語彙サイズ 32k [13] を用いた. 本研究では $T5_{Megagon}$ と表記する. もう一つは, T5 を改良した T5X をベースに, 株式会社レトリバが公開している学習済みの日本語モデル T5 [14] を用いた. 本研究では, $T5_{Retrieva}$ と示す. いずれも事前学習として mC4 と Wikipedia を用いている.

T5 エンコーダへ OCR 誤り訂正候補テキストを入力とし, デコーダー出力を誤り訂正済みテキストとして, 系列変換の枠組みで OCR 誤り訂正の微調整学習を行った. 従って, 入力に誤りがあれば修正された結果を出力し, 誤りがなければ入力をそのまま出力するようにモデルを学習した.

基本的に学習スクリプト [1] を使用した. ハイパーパラメータは, バッチサイズ 32, 最大入出力長 64, 学習率 5e-5, イテレーション 15,000, 勾配累積 2 ステップで, 両モデルとも同じハイパーパラメータを使用した. 定量評価は 3 回の訓練の平均値を示す. また, 事前実験では事前学習済みモデルに対して, 微調整を行わない時の結果はどちらもスコアが 0 となることを確認した.

4.2　ルールによるアプローチ

また, 言語モデルアプローチの性能を比較するために, ルールベースモデルを構築した. 本アプローチは, 国税庁の法人データベースを用いて, 編集距離による検索を行い, 距離が近い文字列があれば, その文字列を修正出力する. 具体的には, データベース [2] を元に企業データベースを作成し, 前処理と重複排除を行い, 約 370 万件のリストを作成した. 前処理はベンチマーク作成時に用いた標準化処理に準じた. 編集距離としてレーベンシュタイン距離を用い, 距離が最小の文字列を候補としてデータベースから検索し, 比較のためのスコアを算出した. スコアは 2 つの部分から構成され, 1 つは編集距離そのものであり,

1) https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py
2) https://www.houjin-bangou.nta.go.jp/

表 5　テストセットにおける言語モデルよる OCR 修正後の結果

Test Set	Method	Ground Truth	Before correction	After correction
Vision	$T5_{Retrieva}$	株式会社 DAISHIZEN	#DAISHIZEN	株式会社 DAISHIZEN
Vision	$T5_{Retrieva}$	東洋エアゾール工業株式会社	東洋工業株式会社	東洋エアゾール工業株式会社
Vision	$T5_{Retrieva}$	株式会社優	株式会社三菱	株式会社三菱ふそう
Vision	$T5_{Retrieva}$	ユニックス	Z	ほっかほっか亭
Vision	$T5_{Retrieva}$	株式会社クリーン・アシスト	株式会社クリーン・アンプト	株式会社クリーン・アップト
Vision	$T5_{Megagon}$	株式会社 DAISHIZEN	#DAISHIZEN	DAISHIZEN
Vision	$T5_{Megagon}$	東洋エアゾール工業株式会社	東洋工業株式会社	東亜工業株式会社
Vision	$T5_{Megagon}$	株式会社優	株式会社三菱	三菱
Vision	$T5_{Megagon}$	ユニックス	Z	会社
Vision	$T5_{Megagon}$	株式会社クリーン・アシスト	株式会社クリーン・アンプト	クリーン・アンプト
Robota	$T5_{Retrieva}$	株式会社 DAISHIZEN	株式会社 DAISHIZEN	株式会社 DAISHIZEN
Robota	$T5_{Retrieva}$	東洋エアゾール工業株式会社	東洋エアゾール工業株式会社社	東洋エアゾール工業株式会社
Robota	$T5_{Retrieva}$	株式会社優	株式会社優	株式会社優
Robota	$T5_{Retrieva}$	ユニックス	ユニックス	ユニックス
Robota	$T5_{Retrieva}$	株式会社クリーン・アシスト	株式会社クリーン・アジスト	株式会社クリーン・アシスト
Robota	$T5_{Megagon}$	株式会社 DAISHIZEN	株式会社 DAISHIZEN	DAISHIZEN
Robota	$T5_{Megagon}$	東洋エアゾール工業株式会社	東洋エアゾール工業株式会社社	エアゾール工業株式会社
Robota	$T5_{Megagon}$	株式会社優	株式会社優	優
Robota	$T5_{Megagon}$	ユニックス	ユニックス	UIX
Robota	$T5_{Megagon}$	株式会社クリーン・アシスト	株式会社クリーン・アジスト	クリーン・アジスト

もう 1 つは編集距離を文字列に対して正規化した比率である. 比率は編集距離を計算したテキストの最大長で割って計算される. 編集距離が 1 以下かつ割合が 0.30 以下の場合, 候補テキストを訂正後テキストとして出力し, そうでない場合は入力テキストが出力される.

5　実験

5.1　定量評価

表 4 に言語モデル T5 とルールベースによる誤り訂正精度を示す. 平均スコアから, $T5_{Retrieva}$ は 94.8% を達成し, 訂正前の精度 85.4%から 9.4 ポイント改善した. 次に, ルールベースの誤り訂正と $T5_{Megagon}$ は, それぞれ 87.2%, 9.3%の精度となった. 同じ T5 とハイパーパラメータを用いても, 事前に学習した重みによって誤り訂正精度が異なることが確認された. 訂正タスクと事前訓練済みモデルの相性や事前訓練済みモデルの学習された時の条件などの違いが要因と考えられる. この実験により, 選択された事前学習済み言語モデルによる誤り訂正が平均的には効果的であり, ルールベースも実用的であることが確認された.

Vision セットは平均結果と同様の傾向を示すが, Robota セットは異なる傾向を示す. $T5_{Retrieva}$ による誤り訂正は, 純粋な OCR 結果よりも精度を向上させたが, 他の方法は精度を悪化させた. OCR の精度が高い場合, 誤り訂正は精度を改善するよりも悪化させる可能性があるため, 実適用時には補正の有無を判断する必要がある.

5.2　定性評価

表 5 にそれぞれのセットにおける T5 モデルの訂正前後及び GT 例を示す. まず, Vision セットの $T5_{Retrieva}$ では「株式会社」が欠損している場合に補正できていることが確認できる一方で, 「Z」から「ほっかほっか亭」と推定し, GT とは異なる補正を行なってしまうことが確認できる. $T5_{Megagon}$ に関しては入力テキストを削除する傾向があり, 不正解になっていることが確認できる. Robota セットは OCR 結果が正しいことが多く, $T5_{Retrieva}$ では補正後も同一の出力ができていることが確認できる. 一方で, $T5_{Megagon}$ では Google セットと同様に入力テキストを削除する傾向がある.

6　おわりに

本論文では, 日本語証憑の OCR 精度を向上させるために, 取引先名における OCR 誤り訂正ベンチマークを構築した. そして, そのベンチマークをもとに, T5 言語モデルによるベースラインを構築し, 2 種類の事前学習済みモデルを用いてその有効性を検証した. さらに, ルールベースのアプローチを用いて言語モデルの有効性を検証した. 実験の結果, 言語モデルアプローチはタスクの内容と事前学習済みモデルの相性がある可能性を示唆し, 効果的な事前学習済みモデルは単純な構造にもかかわらず, OCR 誤り訂正を効果的に実行できることが示された.

謝辞

貴重な議論, 技術的助言, 入念な校正をしてくださった小林一郎教授に感謝する.

参考文献

[1] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In **Proceedings of IEEE Conference on Computer Vision and Pattern Recognition**, pp. 9519–9528, 2022.

[2] Masato Fujitake. A3s: Adversarial learning of semantic representations for scene-text spotting. In **Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing**, pp. 1–5, 2023.

[3] Google. Cloud vision api.

[4] Fast Accounting. Robota api.

[5] Masato Fujitake. Dtrocr: Decoder-only transformer for optical character recognition. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 8025–8035, 2024.

[6] Masato Fujitake. Diffusionstr: Diffusion model for scene text recognition. In **Proceedings of IEEE International Conference on Image Processing**, pp. 1585–1589, 2023.

[7] 竹内孔一, 松本裕治. 統計的言語モデルを用いたocr 誤り訂正システムの構築. 情報処理学会論文誌, Vol. 40, No. 6, pp. 2679–2689, 1999.

[8] 増田勝也. 大域的情報を用いた ocr 文字誤り訂正. 言語処理学会第 21 回年次大会発表論文集, pp. 127–130, 2015.

[9] 謝素春, 松本章代. 日本語 bert モデルによる近代文の誤り訂正. 言語処理学会第 29 回年次大会発表論文集, pp. 1616–1620, 2023.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of International Conference on Neural Information Processing Systems**, pp. 5998–6008, 2017.

[12] Detomo. Japanese_ocr.

[13] megagonlabs. t5-base-japanese-web.

[14] retrieva jp. t5-base-medium.