7ZK-06

近代言語モデルを用いた近代公文書 OCR の精度改善手法の提案

Proposal of a Method for Improving OCR of Modern Official Documents using a Modern Language Model

亀山 京右† 山田 雅之† 中 貴俊† 兼松 篤子†
Keisuke Kameyama Masashi Yamada Takatoshi Naka Atsuko Kanematsu
宮崎 慎也† 長谷川 純一‡
Shinya Miyazaki Junichi Hasegawa

1. はじめに

行政機関に保管されている明治から昭和初期にかけての公文書（近代公文書）は当時の社会情勢や近隣諸国との関係性を知ることのできる貴重な資料である．しかし，近代公文書の多くは旧字，旧仮名遣いを用いた手書き崩し字で書かれており，解読にあたり専門的な知識が不可欠となっている．そのため，利用価値のある資料の多くが活用されていないのが現状である．

我々の研究グループでは，近代の手書き公文書解読を目的とし，近代公文書として体系的に管理されている台湾総督府の公文書を題材とした OCR システム及びデータセットの構築を進めている．これまでの研究開発により，台湾総督府の文書データに対して行画像に書かれている文字及び文字領域の認識精度はおよそ 93%という結果を得られた[1]．しかし，他の公文書データに対する認識率は低く，汎用性に欠けるという問題がある．これを解決する手段として，データセットの拡張が考えられるが，近代文書の手書き文書画像情報，文字情報からなるデータ対を大量に収集することは容易ではない．そのため，汎用性を向上させるためのデータセット拡張は困難となっている．

本研究では青空文庫に公開されている文書データを用いて言語モデルを構築することで近代の汎用的な知識の獲得を目指した．本稿では，近代文語を学習した言語モデルの概要，及び検証結果について述べ，言語モデルを用いた OCR システムいついての検討を述べる．

2. 関連研究

手書き文字の認識を目的とした研究に Minghao らが提案した TrOCR という手法がある[2]．この研究では，英語手書き文字の認識システムを Transformer ベースのモデルを用いて構築している．TrOCR 手法は従来の OCR と以下の点で異なる．

I. 入力画像に対して畳み込み処理を行わず，パッチ分割した画像を用いる．
II. エンコーダのパラメータを事前学習済みモデルで初期化する．
III. デコーダのパラメータの一部を RoBERTa[3]の事前学習済みモデルで初期化する．

本研究では，この TrOCR の構造を参考に，近代手書き公文書解読システムの構築を目的とし，近代文語 RoBERTa モデルの構築を行った．

† 中京大学 Chukyo University
‡ 中京大学人工知能高等研究所
Institute for Advanced Studies in Artificial Intelligence

3. モデルの構築

3.1 データセット

近代文語に対応した RoBERTa モデルの構築のため，青空文庫で公開されている 1950 年以前の近代の書籍 8,004 冊と近代文語コーパス 4 種を用いて RoBERTa 事前学習のためのデータセットを構築した（表 1）．

表 1 事前学習データセット

	データセット	詳細	データサイズ
事前学習	青空文庫	1950年以前の書籍8004冊	305MB
	近代文語コーパス	①明六雑誌コーパス ②国民之友コーパス ③近代女性雑誌コーパス ④NDL digital corpus	①0.78MB ②4.38MB ③2.12MB ④0.06MB

3.2 事前学習

構築したデータセットを訓練・検証用に 8:2 で分割し，モデルの事前学習を行った．RoBERTa のトークナイザとしては BPE が使用されることが多いが，本モデルでは文字単位分割を行うトークナイザを SentencePiece によって，データセットから構築して使用した．SentencePiece は Python で使用可能なライブラリの一つで，データセットから，文字列と ID が 1 対 1 対応した辞書を自動で作成することが可能となっている．文字列の分割単位として，文字，単語，BPE などを選択することができる．

RoBERTa は，入力文字列の一部を[MASK]に置き

換え，その部分のトークンを予測する穴埋めタスク(fill-mask)により訓練を行う MLM(Masked Language Model)となっている(図 1)．レイヤー数を 24，アテンションヘッド数を 16 としたモデル(large-model)を構築し，NVIDIA RTX A6000 を 2 枚搭載した PC を用いて，10 エポック（約 65 時間）学習を行った．

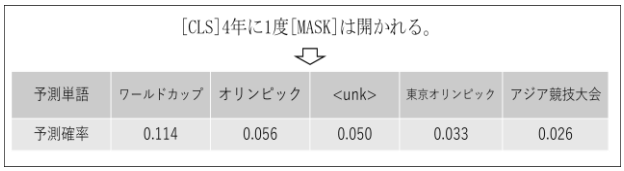

図 1 穴埋めタスク例

3.3 モデルサイズの縮小

事前学習では，学習に用いた文書データの多様性からモデルサイズが小さい状態では安定した学習を行うことができなかった．本研究で構築するモデルは，TrOCR のパラメータ初期化に用いるため，モデルサイズを小さくする必要がある．そのため，レイヤー数を 8 層，アテンションヘッド数を 12 としたモデル(small-model)を新たに構築し，モデルの初期値を large-model の任意のパラメータで初期化することで，学習を安定させることができた．

small-model の学習は large-model と同じデータセットを用いて 5 エポック学習を行った．また，モデルパラメータの初期値は，入力の最初の層と出力前の最終層をそれぞれ対応する層で初期化し，残り 6 層は large-model の 22 層から 6 層を任意に選択して初期化した．

3.4 モデルの検証

モデルの検証データに対する[MASK]予測正解率を表 2 に示す．モデル 2（small-model）との比較のため，small-model と同様のネットワーク構成で large-model のパラメータを用いず，ランダムにパラメータを初期化して学習したモデル 3 を構築した．

モデル 1 の精度は 0.76 なのに対し，モデル 2 の精度は 0.60 となっており，精度の低下を抑えられていることがわかる．また，モデル 3 の精度 0.11 と比べモデル 2 では大きく精度が向上していることが確認できる．そのため，モデルのパラメータ初期値として，学習済みの大きなモデルの任意のパラメータを用いることは，精度改善に大きく貢献していると考えられる．

表 2 各モデルの fill-mask 予測精度

モデル	Accuracy
モデル1 ： large-model	0.76
モデル2 ： small-model (large-modelで初期化)	0.60
モデル3 ： small-model (ランダム初期化)	0.11

4. おわりに

本研究では，TrOCR ネットワークのパラメータ初期化のため，近代文語 RoBERTa モデルの構築を行い，精度の検証を行った．large-model のパラメータを small-model の初期値とすることで，安定した学習を行うことができた．TrOCR では，RoBERTa モデルを知識の蒸留によってサイズ縮小を行っているため，モデルサイズを小さくするために蒸留手法を活用することも効果的だと考えられる．

今回作成したモデルを用いて，近代手書き文字公文書 OCR に向けたモデルのパラメータ初期化を行うことで，データセットの不足を緩和し，認識精度の向上が期待できると考えられる．

謝辞

本研究は JSPS 科研費 JP20H01304，および中京大学戦略的研究「デジタル・ヒューマニティーズプロジェクト:日本近代公文書自動解読システムの開発」の助成を受け，研究，開発を行いました．

参考文献

[1] 山田 雅之，目加田 慶人，長谷川 純一，″歴史的文書データセットの文字矩形情報を用いた行単位画像からの文字列予測と文字セグメンテーション″，情報処理学会論文誌 Vol.65 No.3 1-13 (Mar. 2024)

[2] Minghao et al., ″TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models″,arXiv:2109.10282

[3] Zhuang et al., “A Robustly Optimized BERT Pre-training Approach with Post-training”, CCL (2021)