様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　４　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６７３０１６１

研究課題名（和文）字形情報・言語情報の統合的利用による歴史的文献資料テキスト化システムの高度化

研究課題名（英文）Improvement of Modern Document Textualization System with Integrated Use of Letter Shape Information and Language Model

研究代表者

増田　勝也（Masuda, Katsuya）

東京大学・大学総合教育研究センター・特任助教

研究者番号：２０５１２１１４

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では近代の文献資料に対するデジタルテキスト化の精度向上を目的として、OCR誤り訂正システムの研究開発を行った。デジタルテキスト化の精度評価および精度向上のための言語モデルのための近代の言語リソースを構築し、文字の字形情報と言語情報を組み合わせてOCR誤り箇所の検出、訂正文字候補の生成、訂正文字の選択を行うシステムを構築し、実際に近代書籍のOCR結果に適用し実証実験を行った。OCR誤り訂正の結果をOCRシステムにフィードバックし、OCRシステム自身の精度向上につながることを確認した。

研究成果の概要（英文）：In this research, we have developed an OCR error correction system with the aim to improve the accuracy of digitization of modern documents. We have constructed language resources of modern documents for evaluation of our system and construction of language model for modern documents. We have constructed an error correction system consist of three part, OCR error detection, candidate character generation and selection of a character from candidates. In each part, we use both letter shape information and language model to detect error or to generate candidates. We confirmed that feedback of OCR error correction to the OCR system leads to an improvement of accuracy of the OCR system.

研究分野：　自然言語処理

キーワード：　OCR　デジタルテキスト化　誤り訂正　自然言語処理　デジタルアーカイブ　近代書籍

１．研究開始当初の背景

デジタル・ヒューマニティーズと呼ばれる人文科学とデジタル技術の融合分野が注目を浴びている。この分野においては人文科学系の資料のデジタル化とその利活用が目的の一つとなっており、TEI（Text Encoding Initiative）ガイドラインと呼ばれる人文科学資料のデジタル化・情報アノテーションのガイドラインが策定されてきている。また人文科学資料に対しても、テキストマイニング技術を用いた分析が様々な分野において行われてきており、現時点でデジタルテキストデータが存在しない資料についてもデジタルテキスト化が待たれている。

デジタルテキストを作成する方法としては、人手によるテキスト化が最も高精度に行うことが可能な手段である。しかしながら、人手によるテキスト化は非常に高コストであり、大量に行うことは難しい。また一方で、自動的に行う方法としてコンピュータで資料画像から文字を認識する OCR （Optical Character Recognition）システムを利用する方法がある。OCR システムに関する研究・開発はこれまで多数行われてきており、現代の活字文書に対しては 99%以上の高精度で文字認識が可能である。これらの OCR システムは一般に事前に各文字の特徴を抽出し、それらの特徴量を基に文字間の類似度を求め、与えられた文字画像の文字が何であるかを認識する。そのため、現代の活字に対して構築された OCR システムを近代書籍などの歴史的な文献資料に対して適用すると、同じ文字でありながら字形が現代とは大きく異なる文字が近代書籍には存在するため、それらの文字を誤認識してしまう（表１）。また資料原本の質が悪く、書き込みや汚れ等による認識誤りも発生する。そのため認識精度が大きく低下してしまい、現在の OCR システムをそのまま適用するだけでは、高精度にデジタルテキスト化を行うことができないという問題点がある。これらは主に文字認識に対象文字の画像情報のみを利用していることが原因の一つである。本研究で対象とする画像は近代の書籍をはじめとする文献であるため、書かれている言語としての特徴・情報を利用することでより精度の良い認識を行うことができると考えられる。

表 1：フォントの違いによる誤認識例

画像	誤り例	画像	誤り例
	ご、ざ		戚、咸

２．研究の目的

本研究の目的は、認識対象の字形情報およびテキストの言語情報を統合的に利用した高精度な OCR 文字誤り訂正システムを開発し、特に近代書籍などの歴史的文献資料を対象としたデジタルテキスト化を高精度に行うシステムを開発することである。

本研究では、字形情報と言語情報を用いることにより、文字の見た目の類似度と文字間の言語的つながりの両方の指標を組み合わせて OCR 誤りの訂正を行うことで、高精度にデジタルテキスト化を行うことを目指す。また、言語情報として訂正対象の文字だけではなく対象資料全体の大域的な統計的言語情報もあわせて利用することで、フォントの違い等に起因する、特定の文字が別の文字に誤りやすいなどの誤りパターンを考慮した文字誤り訂正を行うことが可能となる。また、それらのパターンを学習データとして OCR システムにフィードバックし利用することで、OCR システム自身の精度も向上させることが可能となる。

なお本研究の目的は画像から文字を認識・抽出する OCR システムを直接的に作成・改良することではなく、既存の OCR システムの結果に対し字形情報・認識文字の言語情報を利用し、OCR 結果の誤りを高精度に訂正することである。また同時に誤り訂正の際に使用した文字画像と対象文字のパターンを OCR システムに学習データとしてフィードバックし、学習データの充実という形で OCR の認識自体も高精度化する、既存の OCR システムを中心とした統合的な高精度デジタルテキスト化システムの構築を目的とする。

３．研究の方法

（１）近代書籍に対する言語リソースの構築：本研究においての言語モデル用リソースおよび評価用正解データとして、既存の OCR システムを実際の書籍画像データに適用し、それを人手で訂正するという形で近代書籍のテキストデータの構築を行う。

（２）字形情報・言語情報それぞれを用いた OCR 文字誤り訂正システムの構築：字形情報・言語情報を単独で用いた OCR 文字誤り訂正システムを構築する。OCR 誤り訂正システムは①OCR 誤り箇所の検出②訂正文字候補の生成③候補中から訂正文字の選択、という 3 ステップからなり、それぞれについて字形情報、言語情報を用いた処理を行うシステムを構築する。具体的に利用する情報および利用の方法は以下の通りである。

①OCR 誤り箇所の検出

以下の手法で各文字のスコアを求め、ある閾値よりもスコアが低い文字については誤りであると判定する。

・字形情報：OCR システムから出力される確信度を誤り判定のスコアとする。

・局所的言語情報：各文字出現に対しそれを含む文字トライグラム（連続する三文字組）の言語モデル中での出現確率を基にスコアを与える。

・大域的言語情報：OCR 結果中の各文字について、OCR 結果中のその文字の各文字出現を含む文字トライグラムの言語モデルでの出現確率の平均値をスコアとする。

②訂正文字候補の生成

・字形情報１：OCR システムから出力される候補文字を訂正文字候補とする。候補スコアは OCR システムが出力する確信度とする。

・字形情報２：OCR において、同一の文字の OCR 結果の候補として出力されやすい文字を訂正候補とする。候補スコアは OCR 結果において同一の文字の候補文字として出現する確率とする。

・局所的言語情報：各文字出現において、その文字出現を含む文字トライグラムと言語モデルから、その対象文字の箇所に言語的に入るのが尤もらしい文字を訂正候補とする。スコアはその文字トライグラムの出現確率とする。

・大域的言語情報：上記局所的言語情報の出現確率を各文字単位で集計し、入力 OCR 結果全体としてその文字に置き換わるのが尤もらしい文字を候補文字とし、その文字の各文字出現の候補文字とする。スコアはその出現確率の平均とする。

③訂正文字の選択

言語情報を用いて訂正文字の選択を行った。具体的には、候補文字列から辞書を用いて単語列を生成し、どの単語列が言語的に尤もらしいかを言語モデル（単語のトライグラムモデル）を用いて決定する。これにより、訂正後の文字列（単語列）として尤もらしい並びとなるよう候補から訂正後の文字を選択する。

（３）字形情報・言語情報を統合的に用いた OCR 文字誤り訂正システムの構築：上記（２）で開発した各種情報を単独で用いる OCR 誤り訂正システムを組み合わせ、より高精度な OCR 誤り訂正システムを構築する。各ステップにおける組み合わせの方法は以下の通りである。

①誤り箇所の検出

各種情報を用いて計算されたスコアの平均を全体のスコアとし、そのスコアをもとに誤りであるかを判定する。

②訂正文字候補の生成

各手法で生成された候補文字集合を統合し、スコアが上位の文字を全体の訂正文字候補とする。

（４）OCR システムへの誤り訂正結果のフィードバック：上記の OCR 文字誤り訂正システムの結果訂正された文字について、その文字画像と文字のペアを OCR システムに登録し、OCR システム自体で認識可能な近代書籍の文字を増やし、精度を向上させる。

４．研究成果

（１）近代書籍に対する言語リソースの構築：岩波書店「思想」および国立国会図書館内の資料について書籍の画像に対する正解テキストの作成を行った。岩波書店「思想」12 号分、約 2,000 ページおよび国立国会図書館の書籍 20 冊分、約 500 ページについて、既存の OCR システムを実行し、その中の誤り箇所を人手で修正することで正解のデジタルテキストを作成した。

岩波書店「思想」のデジタルテキストについては著作権の関係で一般への公開は難しいが、国立国会図書館の書籍も含め、本研究課題のみならず同様の近代書籍に対するデジタル化の精度向上に関する研究においては今後利用可能であると考えられる。

（２）OCR 誤り訂正システムの構築：前章で示した手法を用いて字形情報・言語情報を組み合わせた OCR 誤り訂正システムを構築した。構築した OCR システムに対し作成した正解データを用いて評価を行った結果、全体としては訂正を行う前よりもテキスト精度（正解テキストに対する一致率）が低下するという実験結果が得られた（表２）。これは、誤り箇所検出において実際には誤りではない文字を誤りとして検出し、それらが後のステップにおいて別の文字に「訂正」されることで「OCR システムが正しく認識していた文字を誤った文字に変更する」という現象が起きたためである。

表２：訂正前後のテキスト精度

データ番号	文字数	テキスト精度		
		訂正前	訂正後	誤り検出を除く
19210010	7746	0.9682	0.9556	0.9762
19400010	8156	0.9935	0.9929	0.9941
19500010	17015	0.9860	0.9441	0.9878
19600010	12866	0.9970	0.9422	0.9972
19700010	28640	0.9713	0.9090	0.9721
19800010	32218	0.9972	0.9842	0.9978
19900010	24379	0.9973	0.9662	0.9976
20000010	15983	0.9977	0.9888	0.9982

そこで、誤り箇所検出が 100%の精度で行われると仮定して、正解データで示されている誤り箇所に対してのみ訂正処理を行ったところ、全体のテキスト精度の向上が見られた（表２「誤り検出を除く」列）。これにより、誤り検出後の OCR 誤り訂正処理（訂正文字候補の生成、訂正文字の選択）については一定の成果が得られていることがわかる。そのため誤り箇所の検出を高精度に行うことができれば、テキスト精度は向上すると考えられる。本研究では誤り箇所検出に字形情報（OCR システムの確信度）と言語情報（対象文字の周辺の文字情報）を用いている。この誤り検出の際には主に「文字の並びとして言語的に尤もらしい」の指標を用いて検出を行うため、既存の近代の言語リソースには出現しない

文字列は基本的には誤りであると検出されてしまう。本研究でも言語リソースの構築を行ってはいるものの、それでも近代の言語リソースは不足している状況であるといえる。

また、本研究で開発した OCR 誤り訂正システムにより訂正された文字の例を表３に示す。これらの文字により、本研究での目的である「字形の違いにより正しく認識できない文字」が実際に訂正さていることがわかる。

表３：訂正された文字の例

訂正前	訂正後	画像
ぐ	く	
逋、適	通	
間、聞	問	

（３）OCR システムへのフィードバックによる OCR システムの精度向上：上記システムで訂正された文字について文字画像を切り出し、その正解文字と文字画像のペアを OCR システムに辞書として登録し、OCR システムがそれらの文字に対応できるように改善を行った。実際に同じ対象資料に対して OCR システムを適用したところ、精度の向上が確認された。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1 件）

① Katsuya Masuda, Makoto Tanji, Hideki Mima. Revealing the Modern History of Japanese Philosophy Using Digitization, Natural Language Processing, and Visualization. Journal of the Japanese Association for Digital Humanities. 査読あり, 1 (1) , pp.37-43, 2015.
DOI: 10.17928/jjadh.1.1_37

〔学会発表〕（計　2 件）

① 増田 勝也. 言語情報と字形情報を用いた近代書籍に対する OCR 誤り訂正. 人文科学とコンピュータシンポジウム 2016, 2016.
http://id.nii.ac.jp/1001/00176186/

② 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, 2015.
http://www.anlp.jp/proceedings/annual_meeting/2015/pdf_dir/P2-2.pdf

６．研究組織

（1）研究代表者

増田　勝也（MASUDA, Katsuya）
東京大学・大学総合教育研究センター・特任助教
研究者番号：20512114

（2）研究分担者

なし

（3）連携研究者

なし