大域的情報を用いたOCR文字誤り訂正

増田 勝也
東京大学 大学総合教育研究センター
masuda@he.u-tokyo.ac.jp

1 はじめに

近年、情報技術の発展に伴い人文科学系分野においてもデジタル・ヒューマニティーズと呼ばれる分野で情報技術の利用が進んでおり、様々な資料のデジタルアーカイブ化についての研究が行われている。書籍のデジタルアーカイブ化に関しては、国会図書館デジタルコレクション[1] をはじめ、Web 上での検索・閲覧が可能なサービスが公開されている。これらのサービスにおいては、書籍のスキャン画像レベルでのデジタル化は進められており、書籍の書誌情報やキーワードによる検索は可能である。しかしながら書籍の内容に踏み込んだ検索やテキストマイニング等を行うためには書籍のテキストが必要となる。膨大な量の書籍に対するテキスト化を人手で行うことは非常に高コストであるため、書籍等活字文書のテキスト化については、OCR(Optical Character Recognition) システムが利用されてきている。OCR による認識は現代の活字文書に対しては 99%以上の精度で行うことが可能であるが、明治期から第二次大戦頃までの近代の活字文書においては現代とのフォントの違い、異体字などの理由により認識精度が低下してしまう。このような OCR 誤りは画像と正解の文字の対応がつけられないことが原因であり、同じ文字については別の同じ文字に誤るという特徴がある。

そこで本論文では大域的情報を用いることで、フォントの違いなどに起因する同じ文字が出現箇所によらず別の同じ文字に誤るような OCR 誤りを訂正することを目的とする。特定の文字の文書中での使われ方を、対象とする文字列全体から集計することでその文字の箇所に入りうる文字の候補を言語モデルを用いて生成し訂正を行う。文字列の各箇所が OCR 誤りであるかどうかは言語モデルを用いて判定し、誤りであると判定された箇所については上記の候補文字によって訂正を行う。実験により、実際にフォントの違いや異体字に起因する認識誤りの訂正が行えることを示す。

[1]http://dl.ndl.go.jp/

表 1: 異体字・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 背景と関連研究

2.1 近代活字文書のデジタルテキスト化

近代の活字文書のテキスト化では、国立国会図書館 NDL ラボ[2]内の翻デジ [2] など、クラウドソーシングを用いた人手による文字起こしが広く行われている。人手によるため精度はほぼ 100%であるが、非常に高コストであり、大量の資料に対して行うことは困難である。一方、OCR システムによる自動テキスト化も近年行われ始めており [4]、低コストでデジタルテキスト化を行うことができるが、近代の活字文書独自の対象文書の状態の悪さ、異体字、現代とのフォントの違いなどの問題点により現代の文書に比べ精度が低下してしまう。異体字、フォントの違いによる誤認識の具体例を 1 に示す。「と」の二画目の入りの点や「感」の「したごごろ」の位置の違いなどにより、現代のフォントで学習した OCR システムでは誤認識してしまう。対象文書の状態の悪さに起因する大規模な誤りは修正することが困難であるが、異体字、フォントの違いによる OCR の誤認識は同じ文字は別の同じ文字として誤って認識されるという特徴があるため、その特徴を考慮し大域的に文字周辺の情報を利用することで、正しい文字の推定が可能であると考えられる。

2.2 OCR 誤り訂正

日本語での OCR 文字誤り訂正手法としては，Noisy channel model により定式化し，文字混同確率モデルおよび言語モデルの確率値の積を最大化する文字列を

[2]http://lab.kn.ndl.go.jp

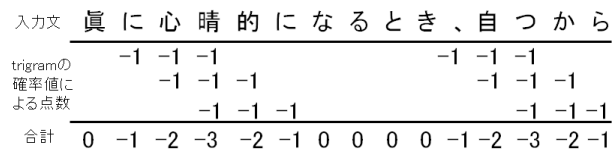

図 1: OCR 誤り箇所の検出例

求める問題とする手法が提案されている．文字混同確率としては文字の図形的特徴を用いたクラスタリングによる文字クラス混同確率を利用する手法 [5] や，文字トライグラムを利用した単語の混同確率を利用する手法 [3] などがあり，言語モデルとしては単語のNグラムが使用されている．後者の研究では紹介論文と同様に，特定分野に特化した文字訂正を行うために，対象データの OCR 結果からNグラムモデルを学習し，訂正文字候補の生成に利用している．これらは文字置換誤りのみに対応可能であるが，近年では文字の融合や分離誤りにも対応する手法 [1] も提案されている．

本研究はこれらの既存研究とは異なり、文字列中の各箇所ごとに訂正候補を作成するのではなく、大域的情報を用いて特定の文字に対し出現箇所によらず訂正候補を生成し、その中で最も可能性の高い文字をその文字の全ての出現箇所での訂正文字とする。各出現箇所における個別の情報はこれらの候補と組み合わせることで精度の高い訂正が可能であると考えられる。

3 提案手法

本研究では OCR システムから出力されたテキストを入力として受け取り、認識誤りを訂正する。一般に OCR 誤り訂正は大きく 1) 誤り箇所の検出、2) 訂正文字候補の生成、3) 候補から訂正文字の選択、の 3 ステップに分けることができる。以下に本手法での各ステップについて詳細を記述する。

3.1 OCR 誤り箇所の検出

各文字の箇所が OCR の誤りであるかどうかの判定は、既存研究 [3] と同様の手法を用いて行う。すなわち、文字 trigram を利用し、その言語モデルにおける確率値が閾値 T_p 以下の trigram が出現した際に、その trigram のすべての文字に対し-1 のスコアを与える。これを文字列の先頭から順に処理していき、最終的にスコアが閾値 T_s 以下の箇所を OCR 誤りであるとする。誤り箇所検出の具体例を図 1 に示す。図 1 では「眞に心晴的になるとき、自つから」という文字列に対し、検出を行っている。下部の「合計」が各箇所のスコアであり、それに基づき誤り箇所であるかどうか判定を行う。なお本論文では $T_p = 0$、$T_s = -3$ とする。すなわち対象とする箇所を含む全ての trigram が言語モデル上に存在しない場合にその対象文字は OCR 誤りであると判定する。

3.2 訂正候補文字の生成

本研究での訂正候補文字の生成は各箇所に対してそれぞれ訂正候補文字を生成するのではなく、文字単位で以下の方法により訂正候補文字を生成する。生成は以下にに示す方法で行う。

1. 各文字 C に対し文字列 $c_1c_2...c_m$ 中の出現位置集合 $P(C) = \{p_i | c_{p_i} = C\}$ を生成する。

2. 文字列中の文字 C を含む trigram の集合 $T(C)$ を生成する。$T(C)$ は各出現位置 $p_i \in P(C)$ に対する以下の trigram を含む集合となる。

$$t_{p_i-2} = (c_{p_i-2}, c_{p_i-1}, c_p)$$
$$t_{p_i-1} = (c_{p_i-1}, c_p, c_{p_i+1})$$
$$t_{p_i} = (c_p, c_{p_i+1}, c_{p_i+2})$$

3. 各 trigram $t_p \in T(C)$ に対し、現在対象とする文字 C の位置に入りうる文字 C_j を trigram t_p における訂正候補として言語モデルから抽出する。その際に以下を trigram t_p における C_j の C に対する訂正候補としてのスコアとして計算する。

$$s(t_p, C, C_j) = \frac{freq(t_p^{C \to C_j})}{\sum_{C_k} freq(t_p^{C \to C_k})}$$

なお、$t_p^{C \to C_j}$ は trigram t_p 中の文字 C を C_j に置き換えた trigram、$freq(t_p)$ は言語モデル中の t_p の頻度とする。

4. 3. で抽出された各訂正候補文字 C_j について、文字列全体における C_j の C に対する訂正候補としてのスコアを以下の式で計算する。

$$S(C, C_j) = \frac{\sum_{t_p \in T(C)} s(t_p, C, C_j)}{|T(C)|}$$

図 2 にアルゴリズムの実行例を示す。例では「戚」という文字に注目し、テキスト中での「戚」を含む trigram を抽出し、各 trigram での「戚」に対する訂正候補とそのスコアを求め、最終的に文字列全体での訂正候補をスコアとともに求める。

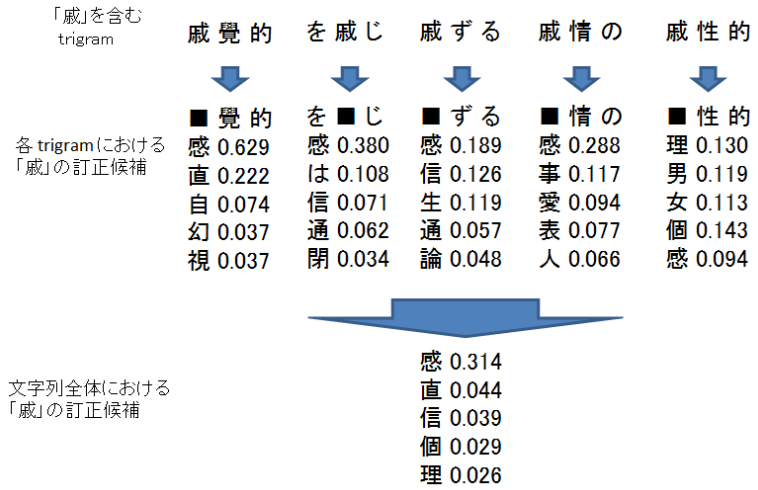

図 2: 訂正文字候補生成例

3.3 訂正文字の選択

前のステップで生成した候補文字から、スコアが最も高い文字 C' を文字 C に対する訂正文字とする。

$$C' = \arg\max_{C_j} S(C, C_j)$$

最初のステップで誤りと判定された文字 C の出現箇所に対し訂正文字 C' で置き換えることにより、訂正後の文字列を生成する。本論文では上記のように非常に単純な選択手法を用いるが、言語モデルを用いて候補文字の中からより尤もらしい文字を選択することで、高精度な誤り訂正を行うことが可能である。

4 実験

本論文では誤り検出および訂正候補の生成に利用する Trigram モデルの学習は青空文庫[3]の全データを用いて行なった。公開されているテキストから本文のみを抽出し、各 Trigram の頻度を求めモデルを構築した。テストには既存研究 [4] でデジタルテキスト化された岩波書店の論文誌「思想」のデータを用いた。「思想」の各年代から 1 論文を抽出し OCR によりデジタル化されたテキストを実験データとし、人手による誤り訂正を行い正解データとして実験を行った。

4.1 実験結果と考察

誤り箇所検出精度

表 2 に誤り箇所検出の精度を示す。各精度は適合率=(検出正解文字数)/(誤りと検出した文字数)、再現率=(検出正解文字数)/(OCR 誤り文字数) により計算する。一部データにおいては誤りであると検出した文字数が非常に多く、適合率が低くなっているが、これは主に文章中の記号や状態の悪さによる大規模な誤認識部分[4]が誤りであると認識されたためである。

表 2: 誤り箇所検出の精度

年代	検出正解数	誤り検出数	誤り文字数	適合率	再現率
1920	358	1338	523	0.267	0.684
1930	177	429	246	0.412	0.719
1940	24	93	53	0.258	0.452
1950	173	1528	239	0.113	0.723
1960	23	612	38	0.037	0.605
1970	661	2089	821	0.316	0.805
1980	63	458	89	0.137	0.707
1990	44	1425	63	0.030	0.698
2000	24	592	37	0.040	0.648

誤り訂正結果

表 3 に実験において実際に訂正された訂正前、訂正後の文字と対象の画像の例を示す。「り」「ル」「と」など本手法で目的としたフォントの違いによる認識誤りの訂正が行えていることが分かる。

表 4 に誤り訂正を行なった前後での文字列全体の精度を示す。精度は精度=(正しい文字数)/(全文字数) により計算する。表中の「誤→正」は誤った OCR 結果が、訂正により正しい文字となった数を示し、「正→誤」はその逆である。「誤→誤」は誤った OCR 結果が訂正により別の誤った文字に変換された数を示す。訂

[3] http://www.aozora.gr.jp/

[4] 本論文ではこの誤認識部分は訂正の対象外である。

表 4: 誤り訂正による精度の変化

年	全文字数	誤り文字数		訂正された文字数			精度	
		訂正前	訂正後	誤→正	正→誤	誤→誤	訂正前	訂正後
1920	9432	523	1338	31	850	320	0.945	0.858
1930	7746	246	459	17	230	160	0.968	0.941
1940	8156	53	90	4	66	20	0.994	0.986
1950	17015	239	1470	26	1299	145	0.986	0.911
1960	12866	38	573	2	550	21	0.997	0.954
1970	28640	821	2039	7	1411	621	0.971	0.922
1980	32218	89	434	12	372	50	0.997	0.986
1990	24379	63	1380	0	1337	43	0.997	0.943
2000	15983	37	573	8	549	16	0.998	0.964

表 3: 訂正された文字の例

訂正前	訂正後	画像
ご	こ	
b	り	
〆、戸	ル	
ピ、霍、芒	と	

正前の誤り文字数に対し、それぞれのテキストにおいて 10%程度の誤りを訂正することができている。ただし、全体としては精度が低下しており、これは「正→誤」の文字数が多いためである。この点については誤り箇所検出の適合率を向上させ、正しい認識であった文字を訂正対象としないようにすることで、文字列自体の精度の低下を防ぐことが可能である。

5 おわりに

本論文ではフォントの違いや異体字に起因する、同じ文字が出現箇所によらず別の同じ文字に誤るような OCR 誤りを訂正する手法を提案した。大域的情報を用いてその文字と置き換えられる可能性が高い文字を求めることで文字単位で訂正のための候補を生成し、文字の訂正を行なった。実験では実際に OCR 誤りの訂正を行い、目的としたフォントの違いによる OCR 誤りが行われたことを示した。実験での全体の文字訂正の精度としては誤り訂正を行うことで精度が低下してしまっているが、これは誤り検出および文字候補の選択に起因することが大きいため、そのステップにおける精度を向上させることで解決することができると考えられる。今後は局所的な情報を利用した OCR 誤り訂正手法と組み合わせることによる、高精度な誤り訂正手法を検討する。また提案手法により生成された訂正文字と対応する画像を OCR システムにフィードバックすることによる、OCR システム自身の精度向上についても検討を行いたい。

謝辞

本研究は JSPS 科研費 26730161 の助成を受けたものです。

参考文献

[1] Graham Neubig, 森信介, 河原達也. 重み付き有限状態トランスデューサーを用いた文字誤り訂正. 言語処理学会第 15 回年次大会 (NLP2009), pp. 332–335, 鳥取, 3 2009.

[2] 永崎研宣. クラウドソーシングによるテクスト翻刻の実践に向けて. 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol. 2014, No. 6, pp. 1–5, may 2014.

[3] 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.

[4] 美馬秀樹, 丹治信, 増田勝也, 太田晋. 近代文献のデジタルアーカイブ化とテキストマイニング-岩波書店「思想」を題材に. 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol. 2012, No. 4, pp. 1–8, 2012.

[5] 永田昌明. 文字類似度と統計的言語モデルを用いた日本語文字認識誤り訂正法. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 98, No. 82, pp. 149–156, 1998.