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

増田　勝也（東京大学 大学総合教育研究センター）

本研究ではOCR結果に対し言語情報および字形情報の両方を利用し近代書籍特有のOCR誤りの訂正を行う．近代の書籍に対するOCR処理では字形の違いにより現代の書籍を対象としたOCRシステムでは一部の文字が正しく認識されない．そこで本研究では近代書籍のOCR結果に対し，そのような誤り傾向を考慮したOCR誤り訂正を提案する．各文字の出現に対し，その周辺の言語情報のみならず対象とするOCR結果全体の中での他の箇所での同じ文字の周辺言語情報を利用することで特定の文字の誤り傾向を考慮し訂正を行う．実験結果により，そのような近代書籍特有のOCR誤りが訂正されたことを示す．

OCR Error Correction for Early-Modern Printed Books by using both Linguistic Information and Character Form Information

Katsuya MASUDA (Center for Research and Development of Higher Education, The University of Tokyo)

This paper describes a OCR error correction method for early-modern books by using both linguistic information and character form information. In OCR for early-modern book, some characters cannot be recognized correctly because of the difference of character form with that in current books. We correct errors in OCR for early-modern books with considering the tendency of errors, which means that a character is misrecognized to the same character in almost all places in the book. In the experimental result, we shows proposed method corrects OCR errors specific to OCR for early-modern books.

１．はじめに

情報技術の発展に伴い様々な資料のデジタル化についての研究が行われている．書籍のデジタル化に関しては，国会図書館がデジタル化した書籍，資料等を国会図書館デジタルコレクションとして公開しているのをはじめ，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による分析等を行うためには機械処理可能なデジタル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，OCRシステムが利用されてきている．OCRによるテキスト化は現代の活字文書に対しては99%以上の精度であると言われているが，例えば近代の書籍においては現代とのフォント・グリフの違い等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．またこのような誤りは画像の情報だけでは正しく認識することは困難であるため，画像認識によるOCRの精度向上ではなく別の種類の情報を用いたテキスト化の精度向上が必要となる．

そこで本論文では言語情報および字形情報の両方を用いることで，近代書籍特有のフォントやグリフの違いに起因するOCR誤りを訂正することを目的とする．特に言語情報については各文字の出現に対する周辺の局所的言語情報のみではなく，それらを文字単位で集計し大域的に言語情報を用いることで，近代書籍でのOCR誤りの特徴である，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誤りの訂正を行う．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，その文字と置き換わるのが尤もらしい文字の候補を，予め学習した言語モデルを用いて生成し訂正を行う．ある文字出現がOCR誤りであるかどうかについても，その文字出現周辺の局所的な言語情報を用いてその文字出現がOCR誤りであるかどうかを判定すると同時に，入力テキスト中のその文字に対する大域的に言語情報を用いて判定する．実験により，実際にフォントの違い等に起因する認識誤りの訂正が行えることを示す．

表 1：フォント・グリフの違いによる誤認識例

画像	誤り例	画像	誤り例
	ご，ざ		戚，成

２．背景と関連研究

近代書籍のテキスト化では，国立国会図書館 NDC ラボ内の翻デジ[1]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や誤りの訂正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，低コストでのデジタルテキスト化が可能であるが，近代書籍独特の対象文書の状態の悪さ，現代の書籍とのフォント・グリフの違いなどの問題点により現代の書籍に比べ精度が低下してしまう．フォント・グリフの違いによる誤認識の具体例を表 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，現代の活字を対象とした OCR システムでは正しく認識できない．このような誤認識は同じ文字は別の同じ文字として誤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言語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．本研究はこれまでの既存研究で行われていた局所的情報を用いた誤り訂正[2]に加え，大域的言語情報や図形情報など様々な情報を組み合わせて訂正を行うことで精度の高い誤り訂正を目指す．

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[3]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[2]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，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[4]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に言語情報を用いて，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 システムから認識結果を入力として受け取り，認識誤りの訂正を行う．OCR システムからは認識結果のテキストのみではなく，各文字の認識の確信度および各文字に対する候補文字も併せて出力されることを前提とする．

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）候補からの訂正文字の選択の 3 ステップに分けることができる．本研究では，各ステップにおいて言語情報，文字の図形的類似情報などを用いて組み合わせて適用することで単一の情報では訂正が困難な誤りについても訂正を行うことが可能となる．なお，以降の記述において「文字」は文字種（「と」や「感」など）のことを表し，「文字出現」はテキスト中のある位置に出現した「文字」のことを指すこととする．

３－１．OCR 誤り箇所の検出

各種情報を用いて各文字出現が OCR 誤りであるかどうかを検出する．検出方法としては以下の方法を組み合わせて用いる．このステップでは OCR システムから出力された OCR 結果文字列を入力とし，各文字出現に OCR 誤りであるかどうかのスコアを与え，そのスコアがある閾値以上である場合にその文字出現を OCR 誤りとする．スコアを与える手法には以下の方法を用いる．

ａ．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．OCR システムが出力する確信度であるため，字形情報をもとにした誤り箇所検出となる．

ｂ．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（連続する三個の文字出現組）の言語モデル中での頻度を基にスコアを与え，それがある閾値未満の文字出現を誤りとする[2]．具体的には以下の手法で判定を行う．なお，入力された OCR 結果の文字列を$c_1c_2 \ldots c_n$（c_i：i番目の文字出現)とする．

1. 各文字出現c_iに対し，その文字出現から連続する三文字出現(トライグラム）$t_i = c_ic_{i+1}c_{i+2}$を取り出す．
2. 予め大量のテキストから学習した言語モデル中での各トライグラムt_iの頻度f_{t_i}を求める．
3. f_{t_i}が閾値T_f 以下の場合，各文字出現c_i, c_{i+1}, c_{i+2}，に対してスコア-1を与える．

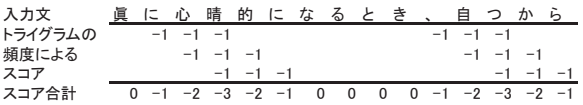

図1：局所的言語情報による文字誤り検出例

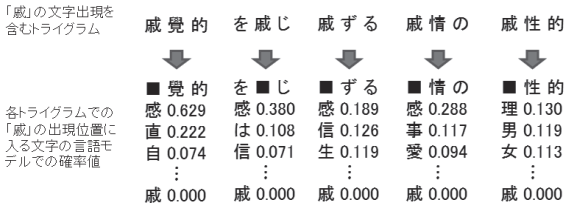

図2：大域的言語情報による文字誤り検出例

4. 全ての文字出現に対して上記の処理を行うことで，各文字出現の誤り判定スコアを決定する．実際の検出例を図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OCR結果を入力とし検出を行う．この例の場合「に心晴」,「心晴的」「晴的に」「，自つ」「自つか」「つから」というトライグラムの出現頻度が閾値以下であり，それらが含む各文字出現にスコア-1が与えられている．それらの合計が最終的な各文字出現の誤り判定スコアとなる．

ｃ．大域的言語情報：入力された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の出現確率からその文字が認識誤りであるかを判定し，誤りと判定された場合にはその文字のテキスト中でのすべての文字出現を誤りとする．具体的には以下の方法で判定を行う．なお，入力された OCR 結果の文字列を$c_1c_2 \dots c_n$（c_i: i番目の文字出現)とする．

1. OCR 結果中の各文字C_jの文字列$c_1c_2 \dots c_n$中の出現位置集合$S(C_j) = \{i|c_i = C_j\}$を生成する．

2. $S(C_j)$から文字列中で文字C_jを含むトライグラムの集合$T(C_j)$を生成する．

$$T(C_j) = \{t_i|t_i = c_{i-1}c_ic_{i+1}, i\epsilon S(C_j)\}$$

3.各トライグラム$t_i \in T(C_j)$に対し，対象としている文字C_jがその位置に入る確率を以下の式で求める．

$$P(C_j, t_i) = \frac{freq(t_i)}{\Sigma_{C_k} freq\left(t_i^{C_j \to C_k}\right)}$$

$t_i^{C_j \to C_k}$は，トライグラムt_i中の文字C_jをC_kに置き換えたトライグラムを表し，$freq(t)$はトライグラムtの言語モデル中での出現頻度とする．この式では，与えられたトライグラムt_iにおいて，文字C_j以外の文字が決まったときにC_jが出現する確率を表す．

4. 文字C_jの誤り判定スコアを 3.で求めた確率値の平均とする．すなわち以下の式で表される．

$$Score(C_j) = \frac{\Sigma_{t_i \in T(C_j)} P(C_j, t_i)}{|T(C_j)|}$$

5. C_jの各文字出現の誤り判定スコアを 4.で計算された文字C_jの誤り判定スコアとする．

実際の検出例を図2に示す．この例では「戚」という文字を対象とし，入力されたOCR結果の文字列においてこの文字が誤りであるかどうかを判定する．まず入力文字列中の「戚」の文字出現をすべて抽出し，それを含むトライグラム集合（「戚覺的」「を戚じ」「戚ずる」「戚情の」「戚性的」）を生成する．その各トライグラムの言語モデル上での確率を求め，それを平均した値を誤り判定スコアとする．この場合はすべてのトライグラムについて言語モデル上での確率値が0である（すなわち言語モデル上ではこのトライグラムは出現しない)ため，誤り判定スコアは0となる．

局所的言語情報での誤り判定は文字出現に対して行うのに対し，大域的言語情報ではある文字のすべての文字出現に同一の誤り判定スコアを与える．これは，今回の主な対象である「字形の違いにより同一の文字が別の同一の文字に誤る」というOCR誤りに対応するものである．

以上の三種類の情報を用いた誤り判定スコアにより各文字出現がOCR誤りであるかどうかを判定する．各手法を組み合わせる際には，各手法で出力される誤り判定スコアをスコアが高いほど誤りとなるよう正規化した上で平均しそれが閾値以上の場合には誤りであるとする．

３－２．訂正文字候補の生成

前ステップで誤りであると判定された文字について，その文字に置き換わる訂正文字の候補を生成する．訂正文字候補生成においては併せて訂正候補としてふさわしいかどうかを表すスコアを付与する．方法としては以下の方法を用いる．

ａ．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．

ｂ．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．候補の生成方法は以下の方法で行う．

1. あらかじめ行なった大量のOCR結果から，各文字出現におけるOCR結果とその候補文字をあわせた集合$S_i = \{C_1, C_2, \dots, C_n\}$を作成する．

2. 文字C_iに対するC_kの候補文字スコアを以下で定義する．

$$Score(C_k, C_i) = \frac{|\{S_j|S_j \ni C_i \wedge S_j \ni C_k\}|}{|\{S_j|S_j \ni C_i\}|}$$

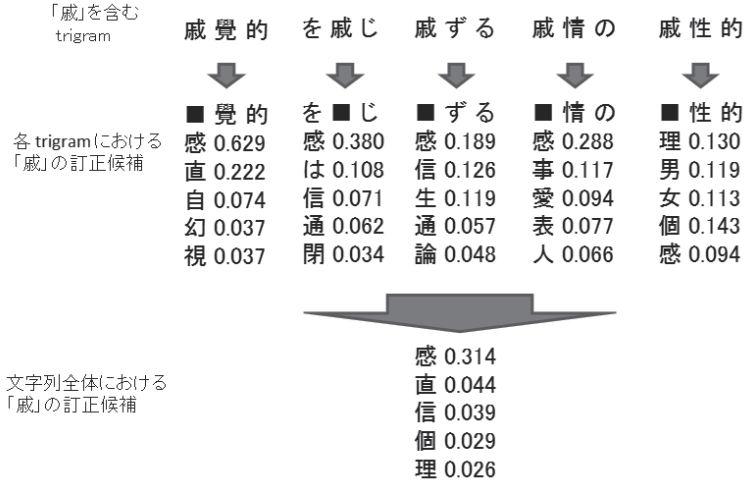

図３：大域的言語情報による誤り文字候補生成

すなわち，OCR結果またはその候補文字としてC_iが含まれる文字出現のうち，C_kが含まれる文字出現の割合をスコアとする．これは，OCR システムにおいて利用されている画像上の文字の字形情報の類似度を OCR 結果から推定しているスコアとなる．本来は画像上での字形の類似度を直接的に計算すべきであるが，すべての文字についてそれらを計算するのは高コストであるため，OCR 結果から推定を行っている．

ｃ．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．具体的には以下の方法で求める．

1. 各文字出現c_i（文字 C)に対しその文字を含む以下のトライグラムを抽出する．

$$t_{i-2} = c_{i-2}c_{i-1}c_i$$
$$t_{i-1} = c_{i-1}c_ic_{i+1}$$
$$t_i = c_ic_{i+1}c_{i+2}$$

2. 上記の各トライグラム t_j に対し文字 C に置き換わるべき候補文字を言語モデルから求める．各トライグラムについて，以下の確率値が高い文字C_kを言語モデルを用いて求める．

$$P(C, C_k, t_j) = \frac{\mathrm{freq}\left(t_j^{C\to C_k}\right)}{\Sigma_{C_l}\mathrm{freq}\left(t_j^{C\to C_l}\right)}$$

$t_j^{C\to C_k}$はトライグラムt_j中の文字CをC_kに置き換えたトライグラムを表し，freq(t)はトライグラムtの言語モデル中での出現頻度とする．

3. 各トライグラムの確率値の平均を取りその平均確率値の高い文字を候補文字とする．その際のスコアは平均確率値とする．

ｄ．大域的言語情報：ｃ．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[5]．具体的には以下の方法で行う．

1. c.の局所的言語情報を利用した手法により各文字出現c_iに対し，候補文字集合Q_iを作成する．

2. OCR 結果中の各文字C_jの文字列$c_1c_2 \ldots c_n$中の出現位置集合$S(C_j) = \{i|c_i = C_j\}$を生成する．

3. 各出現位置$i \in S(C_j)$の候補文字集合Q_iから文字C_jに対する候補文字集合を生成する．候補文字はQ_iに含まれるすべての文字とし，その候補文字スコアは全てのQ_iでのスコアの平均とする．

具体例を図３に示す．図３の例では，「戚」という文字に対する訂正候補文字を生成している．入力された OCR 結果テキストから，「戚」を含むトライグラムを全て抽出し，「戚」の各文字出現位置に入るべき文字候補を言語モデルでの確率値により求める．そして文字「戚」の訂正文字候補としては，５箇所の「戚」の文字出現に対して求めた訂正候補の確率値を文字ごとに平均し，その平均値の高い順で「感」「直」「信」「個」「理」が訂正文字候補として求められている．

以上の各手法により生成された候補文字を用いて最終的に訂正文字の決定を行う．各種法で生成された候補文字を組み合わせて使用する場合には，それらのスコアを正規化した上で和を取り候補集合をマージすることでひとつの候補集合を生成する．その際には，スコア順で上位の文字のみを次のステップで使用する候補文字とする．

表２：誤り検出結果

データ番号	文字数	誤り文字数	誤り検出数	誤り検出正解数	誤り検出正解率	誤り検出再現率
19210010	7746	246	403	189	0.4690	0.7683
19400010	8156	53	53	29	0.5472	0.5472
19500010	17015	239	1300	155	0.1192	0.6485
19600010	12866	38	819	26	0.0317	0.6842
19700010	28640	821	2603	671	0.2578	0.8173
19800010	32218	89	493	60	0.1217	0.6742
19900010	24379	63	1173	44	0.0375	0.6984
20000010	15983	37	298	27	0.0906	0.7297

表３：誤り訂正結果

データ番号	文字数	誤り検出数	訂正文字数	誤り文字訂正数	訂正正解数	テキスト精度		
						訂正前	訂正後	誤り検出を除く
19210010	7746	403	269	128	43	0.9682	0.9556	0.9762
19400010	8156	53	26	18	3	0.9935	0.9929	0.9941
19500010	17015	1300	842	110	20	0.9860	0.9441	0.9878
19600010	12866	819	731	25	0	0.9970	0.9422	0.9972
19700010	28640	2603	2371	574	12	0.9713	0.9090	0.9721
19800010	32218	493	371	43	7	0.9972	0.9842	0.9978
19900010	24379	1173	795	34	0	0.9973	0.9662	0.9976
20000010	15983	298	164	17	5	0.9977	0.9888	0.9982

３－３．候補から訂正文字の選択

前のステップで生成された訂正文字候補の中から最終的に訂正結果として出力する文字を選択する．方法としては[2]で提案されている手法と同様の手法を用いる．すなわち，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．

４．実験

実験データには[6]で使用されている岩波書店「思想」のデジタル化データを用いた．OCRシステムにはメディアドライブ社のWinReader PRO を使用し，訂正文字候補及びその確信度を出力できるようカスタマイズを行なった．OCR誤り訂正の正解データは，OCRシステムから出力されたOCR結果のテキストに対し人手で修正を行ない作成した．また，訂正に使用する言語モデル（文字トライグラムモデル，単語トライグラムモデル）および単語辞書は，青空文庫にて公開されている書籍の本文データを利用し作成した．単語辞書を作成する際には，形態素解析器MeCab[7]を利用し単語分割を行った．またOCR結果を用いた文字の図形的類似情報は[6]で行われた岩波書店「思想」のOCR結果を用いて計算を行った．

まず誤り検出の実験結果を表２に示す．実験データの番号は，対象としている岩波書店「思想」の出版年と論文番号を表している．今回は各年代において一論文を抽出し実験を行なった．「文字数」は提案手法で入力として受け取るOCR結果の全文字数を表す．「誤り文字数」はOCR結果の時点で含まれるOCR誤りの文字数，「誤り検出数」は提案手法でOCR誤りとして検出された文字の総数，「誤り検出正解数」は「誤り検出数」のうち，実際にOCR誤りである文字数を表す．「正解率」および「再現率」は以下の式により計算される値である．

$$\text{正解率} = \frac{\text{誤り検出正解数}}{\text{誤り検出数}}$$

$$\text{再現率} = \frac{\text{誤り検出正解数}}{\text{誤り文字数}}$$

誤り検出の正解率は年代によりかなりばらつきがあることがわかる．一部年代のデータにおいては，誤りであると検出された文字数が非常に多く「正解率」が非常に低くなってしまっている．これは主に文章中の記号や原本の状態の悪さに起因する大規模な誤認識部分が含まれているからである．これらの誤認識部分は今回の提案手法の対象外であると考えられるが実際の誤り訂正と

表４：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	

しては何らかの対応が必要である．一方，再現率に関してはどの年代においても 60～70%となっており，OCR 誤りの半数以上の文字を誤りであると検出できている．

また，誤り検出結果を含めた OCR 誤り訂正結果を表３に示す．「誤り検出数」は表２と同じ値である．「訂正文字数」は誤りであると認識された文字のうち提案手法により実際に別の文字に訂正処理が行われた文字数である．「誤り文字訂正数」は，「訂正文字数」のうち OCR 結果の時点で OCR 誤りであった文字数を表す．「訂正正解数」は「誤り文字訂正数」のうち訂正処理を行なったことで正しい文字に変換された文字数を表す．「テキスト精度」は「文字数」に対し正しく認識されている文字数の割合を示す．すなわち，「訂正前」の「テキスト精度」は OCR システム自身の精度，「訂正後」は提案手法での OCR 誤り訂正後の対象テキスト全体の精度を表す．

実験結果からは今回の提案手法においては OCR 誤り訂正を行うことで全体のテキスト精度としては訂正前よりも悪くなっていしまっていることが見受けられる．精度の悪化については誤り検出の精度の低さが主な要因である．誤り検出の精度が低いため，OCR 結果の時点で正しい文字が誤りであると認識され，その後の訂正処理により正しい文字から誤った文字への訂正がおこなわれてしまい，全体として精度が悪化してしまっている．実際に，誤り検出の精度が 100%であると仮定して，正解データの誤り箇所のみの訂正を行うと，多いものではテキスト精度として 1%弱の精度向上が見られた．

表４に実際に正しく訂正が行われた文字の例を画像とともに示す．本論文での主な対象である，「字形の違いにより正しく認識でない文字」が実際に訂正されていることが分かる．

５．まとめと今後の課題

本論文では複数種類の情報を用いて OCR 誤りの訂正を行う手法を提案した．言語情報および字形情報の両方を用いることで，それぞれ単独では訂正が困難な文字についても訂正を行うことが可能となった．

今後の課題としては，誤り箇所検出の精度向上が第一に挙げられる．現状では誤り箇所検出の精度が低いため，本来訂正する必要のない文字についても訂正処理が行われ，正しく認識された文字を誤った文字に訂正してしまう現象が起こっている．また，現状は異なる種類の情報を，その情報を利用した誤り検出・候補生成の結果をスコア化し単純に和を取ることで組み合わせている．しかしながら，組合せの別の計算方法やスコア化以外の組合せ手法も考えられるため，それらの検討も必要であると考えられる．

また今後は，今回の手法により抽出された画像と文字の対応を OCR システムにフィードバックし，学習データとして利用することで OCR システム自体の精度向上に繋がるか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) 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（2014）.
2) 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）.
3) 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）.
4) 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 回年次大会発表論文集, pp. 332-335,（2009）.
5) 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）.
6) 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）.
7) Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004).