自然言語処理 130－3
（1999．3．4）

統計的形態素解析と文字n-gramを利用したOCR誤り訂正

竹内孔一† 松本裕治‡
†学術情報センター 研究開発部 ‡奈良先端科学技術大学院大学 情報科学研究科
〒112 東京都文京区大塚3－29－1
E-mail: koichi@rd.nacsis.ac.jp

近年，インターネットの普及により，OCRを用いたテキストの電子化がますます重要な課題となってきた．日本語OCR誤り訂正の先行研究において統計的言語モデルを利用して訂正対象と同じ分野の学習コーパスを用意することで高精度の訂正能力を示す研究がある．しかし，電子化された大量テキストコーパスを期待できない場合が多い．そこで電子化されたコーパスがない分野に対して，OCR処理された誤りを含むテキストから学習を行なうモデルを構築する．この時，辞書に無い未知語獲得もOCR処理されたテキストから行なう．実際にOCR処理されたテキストに対する訂正実験の結果，学習コーパスと訂正対象の分野が一致していた先行研究に比べ約1/4程度の訂正精度を示したことを報告する．

キーワード 統計的形態素解析，OCR誤り訂正，文字n-gram

OCR Error Correction Using Stochastic Morphological Analyzer with Character N-gram Model

Kouichi Takeuchi † Yuji Matsumoto ‡
† National Center for Science Information Systems
‡ Graduate School of Information Science,
Nara Institute of Science and Technology

In recent years, OCR error correction is becoming more and more important technique for the purpose of converting printed texts into electronic ones on computers. As a previous work, there are some studies of OCR post processor which show high performance of error correction when they use a large on-line corpus which is the same domain as their target of correction. However, we cannot prepare large on-line corpus at every domain. In this paper, we present an OCR error correction method which uses OCR's output texts in a domain in which no large scale training text exists. We also show some methods to get unknown words using OCR's output texts. When our method is applied to error correction of OCR's output texts, the experimental results shows that the performance is quarter as much as our previous result in which target text and a on-line corpus are of the same domain.

key words stochastic morphological analysis, OCR error correction, character n-gram

1 はじめに

近年，インターネットの普及により電子化された情報は世界中の人々が利用できる環境となりつつある．その中で文字情報は基本媒体であることから，文字情報の電子化は重要な課題である．印刷された文字を読み取るにはＯＣＲ文字読み取り装置を利用する．ＯＣＲは文字に関する画像情報をもとに画像から文字列に変換するが，その結果には言語的に見て明らかに誤りとなる文字列が含まれる．そこで，本論文では言語的な情報を用いてこれらの誤りをどの程度訂正できるかを明らかにしたい．

英語におけるスペルチェックと訂正の技術は進んでおり，Kukich[3] にまとめられている．英語は日本語と異なり，単語ごとに区切りがあるため，単語単位で解析を行うことができる利点がある [8]. Tong[7] らは OCR の文字誤り訂正において，単語間の文字の異なりから単語間の距離を推測し，単語 bigram 確率モデルで候補を選択するモデルを提案し効果を上げている．しかし日本語では単語間の区切りが明示されていないためこれらの方法を直接利用できない．

日本語に対して単語のわかち書きの曖昧性も考慮した複数の単語候補の中から動的に最適な単語列を選択する手法 [12][9][11] が提案されている．永田 [9] は，文字間の画像類似度と統計的な言語モデルを用いて EDR を用いた疑似誤りデータに対する訂正実験を行ない高い精度を得ている．また画像的な類似度を用いない手法として，我々が提案した手法 [11] は統計的な言語モデルのみを利用して新聞記事テキストのランダム誤りデータに対する訂正実験の結果高い訂正精度を得た．ただし，これらの高い精度の獲得には言語モデルの学習に用いたコーパスと解析対象のテキストが同じ分野である必要[1] がある．現在テキストコーパスの拡充は進んでいるが，分野によっては学習に使用できる電子化されたコーパスが存在しない場合も多い．永田 [9] は学習に使用したコーパスと異なる分野の実際の OCR 誤り結果に対して訂正実験を行なったが，未知語の部分では訂正がうまく働かなかったことを報告している．

本論文では既に OCR 処理されたテキストから文字の連接確率や未知語を獲得する手法を示し，コーパスの無い分野に対して訂正するモデルを提案する．つまり，他分野であるが大量テキストデータで獲得された統計量とＯＣＲ処理された誤りを含むテキストから統計量を融合してモデルを構築する．訂正システムの基本的な構成は [11] と同様であるが学習するコーパスと未知語獲得の部分が異なる．学習コーパスとの関係も含めて OCR 誤り訂正モデルの全体図を図１に示す．文字誤り訂正と候補の生成には文字 trigram を利用し，候補の選択には統計的形態素解析システム（品詞 trigram モデル）を利用した．未知語の獲得には品詞 trigram モデルと n-gram を利用した．文字間の画像的な類似度，および OCR の文字候補は利用しない[2] ．

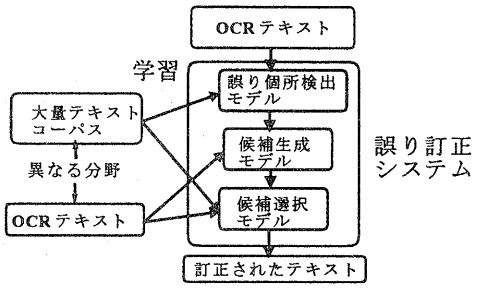

図 1: OCR 誤り訂正システム

実験は実際のＯＣＲ誤りデータに対して訂正を行う．データは奈良先端科学技術大学院大学電子化図書館で蓄積しているＯＣＲ処理されたバイオサイエンス関連の論文誌を用いた．誤りを含むテキストデータからの学習によっても訂正が可能である事を示し，[11] の場合に比べて約 1/4 の訂正能力を示したことを報告する．

2 ＯＣＲ誤り訂正システム

OCRの誤り訂正は次のように定式化される．OCR の入力文字列を $C = \{c_1, c_2, ..., c_m\}$ とし，その出力文字列を $X = \{x_1, x_2, ..., x_k\}$ とする．訂正により最適な文字列 $\hat{C}$ を選択するために訂正モデル $P(C|X)$ の確率値を最大化する．

$$\hat{C} = \arg\max_{C} P(C|X) \quad (1)$$

ベイズの定理から

$$= \arg\max_{C} \frac{P(X|C)P(C)}{P(X)}$$

[1] 学習コーパスの分野の異なりによる影響について [5][10] に実験結果がある．

[2] これは既に OCR 処理されたテキストをあとから訂正する場合，現実的に起こりえる状況である．

$$= \arg\max_{C} P(X|C)P(C) \quad (2)$$

となる．$P(X|C)$ は OCR 装置の確率モデルである．我々の提案するシステムは文字に関する画像情報を用いないため，この確率を 文字 trigram による前後文字列の確率から推定する．$P(C)$ は言語モデルを示している．訂正システムは以下の 3 つの手順で構成されている．

1 文字誤り箇所の検出 (Detection)．

2 辞書引きによる単語候補の生成 (Generation)．

3 候補の選択 (Selection)．

2 において単語候補を生成するので，誤り訂正は最適な単語列$\hat{W}$を選択することになる．(2) 式を書き直して

$$\hat{W} = \arg\max_{W} P(S|W)P(W) \quad (3)$$

となる．ここで，S は OCR 出力の単語列，W は単語候補列である．$P(W)$ は言語モデルで，候補の選択で導入する．$P(S|W)$ はこれら単語列の変換確率を示しており以下の式で推定する．

$$P(S|W) = \prod_{i=1}^{n} P(s_i|w_i) \quad (4)$$

n は文中の単語数を示す．s_i と w_i はそれぞれ OCR の出力単語，候補の単語であり，$P(s|w)$ が単語の変換確率である．この式は単語候補生成の部分 (2.2.2 節) で計算される．

従来の方法 [6] では誤り訂正システムの言語モデルの学習に使用するテキストコーパスと訂正対象のテキストとは同じ分野としていた．しかし，訂正対象のテキストと学習に用いる大量コーパスとの分野が異なるため，訂正対象のテキストそのものも学習に組み込む．よって学習コーパスとして

a. 大量にあるテキストデータ

b. 誤り訂正を行なう OCR 処理されたテキストデータ (以降 OCR テキストと呼ぶ)

の 2 つを仮定する．**a** と **b** は異なる分野のデータとする．使用する言語モデルによって学習コーパスは変更して用いる (図 1 参照)．以下，**1** 誤り箇所の検出，**2** 候補の生成，**3** 候補の選択について各モデルとその学習方法を示す．

2.1 文字誤り箇所の検出 (Detection)

●使用する言語モデル

文字 trigram 確率を用いて文字誤り検出を行う．文字列 c において確率 $P(c_i|c_{i-2}, c_{i-1})$ が T(足切り値) 以下なら c_iのみを誤り文字とするのではなく，c_{i-1}や c_{i-2}も誤り文字の対象とする．そこで文字 trigram 確率が Tp 以下である 3 文字列に対して各々−1 点を与え，これを文字列の文頭から順に当てはめて行き，各文字において Ts 以下の点のついた文字を誤りとする．ここで $Tp = 0$, $Ts = -2$ とした ([6] 参照)．

●学習について

誤り箇所検出の文字 trigram のパラメータ推定にはＯＣＲテキストの分野と異なる **a** 大量テキストから行なう．これは誤り箇所検出で指摘されなかった文字は訂正されないため指摘洩れを少なくするためである．当然，分野で異なる専門用語の部分では誤りを多く指摘するが文字候補モデルを OCR テキストから学習させることで補う．

2.2 単語候補の生成 (Generation)

文字 trigram モデルを利用して候補文字を作成する．想起した文字を元に辞書引きを行い単語候補を生成する．その際，文字候補の順位から各単語に対して変換確率を計算する．

2.2.1 文字列候補の生成

●使用する言語モデル

文字 trigram モデルを用いて文字候補の生成を行う．最大で 2 文字連続まで候補を出力する．以下は 2 文字列の候補文字の例である．誤り検出部で指摘された文字を m_i, m_{i+1} とし [11][6]，その前後の文字列を $c_{i-2}, c_{i-1}, .., c_{i+2}, c_{i+3}$ とする．まず，文頭からの以下の確率 P_fを計算し，m_i, m_{i+1} それぞれに対して出現確率の上位 5 個の文字候補を推定する．

$$\begin{aligned} &P_f(c_{i-2}, c_{i-1}, m_i, m_{i+1}, c_{i+2}, c_{i+3}) \\ &= P(m_i|c_{i-2}, c_{i-1})P(m_{i+1}|c_{i-1}, m_i) \\ &\quad \times P(c_{i+2}|m_i, m_{i+1})P(c_{i+3}|m_{i+1}, c_{i+2}) \end{aligned} \quad (5)$$

次に，文末方向から同様に以下の確率

$$\begin{aligned} &P_b(c_{i-2}, c_{i-1}, m_i, m_{i+1}, c_{i+2}, c_{i+3}) \\ &= P(m_{i+1}|c_{i+2}, c_{i+3})P(c_{i-1}|m_i, m_{i+1}) \\ &\quad \times P(c_{i-1}|m_i, m_{i+1})P(c_{i-2}|c_{i-1}, m_i) \end{aligned} \quad (6)$$

を計算して上位5個の文字列を加えて10候補作成する．

●学習について

候補文字を生成する文字 trigram モデルはOCRテキストから学習させる．これにより候補は OCR テキストに沿った文字候補を生成する．ただし，OCR テキストは1）誤りを含むこと，2）量が少ないことから以下のようにした．

1. 文字 trigram 頻度が低い接続は確率計算に含めない．
2. 未出現文字に対して最低確率値を設ける．

1. ではしきい値 (T_k= 4) 以下の頻度は切り捨てた．これは獲得された文字 trigram において低頻度 (1 ~ 3) な文字列に誤り接続が多く含まれていたので設定した．2. の最低確率値は OCR テキスト中に出現した全文字種 N_kを測り，$1/N_k$とした．よって候補文字の文字 trigram 確率は

$$P_{ocr}(c_i|c_{i-2},c_{i-1}) = \begin{cases} \frac{C(c_i,c_{i-2},c_{i-1})}{C(c_{i-2},c_{i-1})}(1-\frac{1}{N_k}) & C(c_i,c_{i-2},c_{i-1}) > T_k \\ \frac{1}{N_k} & C(c_i,c_{i-2},c_{i-1}) \leq T_k \end{cases}$$

となる．ここで $C(\cdot)$ はコーパス中の文字列の頻度を表す．これを 2.2.1 節の (5) (6) 式に当てはめて，同様に文字候補を生成する．

2.2.2 辞書引きと単語変換確率の計算

上記の手順により獲得された候補文字から辞書引きを行い，単語列を生成する．辞書は統計的形態素解析システム (ChaSen)[13] の辞書 (約 17 万語) を利用した．辞書になく辞書引きできない部分の文字列がある場合は未知語として文字列を出力する

辞書引きを終えた後，文字 trigram による文字候補を利用して単語の変換確率を計算する．候補単語 w_iの各文字を l とし，$L(w_i)$ を単語 w_i の文字集合とする．s_iを単語 w_iに対応する入力文の文字列 (単語) として，単語の変換確率 $P(s_i|w_i)$ は以下のように近似した．

$$P(s_i|w_i) \approx \prod_{l \in L(w_i)} \alpha^{k_l-1}\beta^h \qquad (7)$$

ここで，文字 l は文字候補において上位から k_l 番目であり，単語 w_iは単語 s_i から h 個文字が入れ替わっているとした．つまり，α^{k_l-1} は単語 w_i の信頼度をあらわし，β^hは入力文字列と候補単語との距離 (edit distance) をあらわしている．

αとβは実験的に人手で決定した．αとβは同じ値を用いており実験では 0.0001，0.0005，0.001 の値を用いた．

2.3 候補の選択 (Selection)

●使用する言語モデル

(3) 式で示したように誤り訂正モデルは言語モデル $P(W)$ を必要とする．本論文では候補選択のための言語モデル $P(W)$ として統計的形態素解析システム（品詞 trigram モデル）を使用した．

$$P(W) \approx P(W,T) = \prod_{i=1}^{n+2} P(w_i|t_i)P(t_i|t_{i-2},t_{i-1}) \qquad (8)$$

ここで W，S はそれぞれ単語列，品詞列を表している．また w は単語，s は品詞を表している．t_{-1}，t_0は文頭，t_{n+1}は文末，w_{n+1}は空語を示している．

●学習について

品詞 trigram モデルはパラメータとして品詞の接続確率 $P(t_i|\ t_{i-2},\ t_{i-1})$ と単語の生成確率 $P(w_i|\ t_i)$ がある．これらのパラメータを推定するためには大量なコーパスが必要となるが解析対象である OCR テキストと同分野の大量コーパスは存在しないと仮定している．そこで，品詞の接続確率は分野が異なっても分野による影響が少ないと考え大量なテキストコーパスを利用し，単語の生成確率は OCR テキストから学習する．

品詞接続確率を推定するには，テキストコーパスを形態素解析システム[3] で解析し，その出力から品詞接続頻度を数え上げることで獲得する．

一方，単語の生成確率の獲得において OCR テキストから辞書にない未知語を獲得し確率を付与することは容易ではない．そこで，以下の2通りの方法について提案する．

(イ) 形態素解析システム[3] を利用して未知語を抽出し確率を付与する．
(ロ) 未知語獲得方法として文字 n-gram を用いて (イ) に埋め込む．

これらの各々の場合について単語の生成確率の推定方法を以下で説明する．

[3] ここでは文献 [11] の新聞記事で学習させた品詞 trigram モデルを利用した．

2.3.1 (イ) 形態素解析システムを利用した単語生成確率と未知語の抽出

形態素解析システム[3]は辞書に無い文字列が入力されると未知語として文字列を出力する．よってOCRテキストを形態素解析システムで解析した結果から単語の頻度を数え上げて確率 $P(w|t)$[4] を獲得する．形態素解析システムが未知語として出力する単語はサ変名詞として他の辞書にある語と同様に数え上げる．ただし，すべての単語に対して，低頻度 (しきい値 T_t=4 以下) の単語は捨てる．捨てた単語は未知語のままなので，これらを集計して未知語に与える確率の総和とし，$P(w|$ サ変名詞) から引いておく[5]．つまり，サ変名詞以外の確率は以下のように付与される．

$$P(w|t) = \begin{cases} \frac{C(w,t))}{C(t)} & C(w,t) > T_t \\ P(w|\text{サ変名詞}) & C(w,t) \leq T_t \end{cases}$$

未知語に配分する確率の総和 P_{unk} を低頻度 ($C(w,t) \leq T_t$) の単語から以下のように推定した．

$$P_{unk} = \frac{\sum_{C(w,t)\leq T_t} 1}{\sum_{C(w,t)>0} C(w,t)}$$

ここで $\sum_{C(w,t)\leq T_t} 1$ は頻度 1 から 4 で出現した単語を全て頻度 1 とみなして数え上げている．これを全体の総和 $\sum_{C(w,t)>0} C(w,t)$ で割る．これは頻度 1 の出現回数で頻度 0 の回数を予測するグッド・チューリング法をもとにしている．よってサ変名詞の場合は以下のようになる．

$$P(w|\text{サ変名詞}) = \begin{cases} \frac{C(w,\text{サ変名詞}))}{C(\text{サ変名詞})}(1-P_{unk}) & C(w,\text{サ変名詞}) > T_t \\ \frac{P_{unk}Leng(w)}{\sum_{C(w,t)\leq T_t} 1} & C(w,t) \leq T_t \end{cases}$$

上式において未知語に対して長さによる単語の頻度関数 $Leng(w)$ をかけた．文献 [4] ではポワソン分布を利用しているが，我々はOCRテキストから頻度を利用して獲得した[6]．

$$Leng(w) = \frac{\sum_{length(w)} C(w, length(w))}{\sum_{C(w,t)>0} C(w,t)} \quad (9)$$

ここで $length(w)$ は単語 w の長さを返す関数である．$C(w, length(w))$ は長さ $length(w)$ の単語の頻度を表している．

[4] (8) 式中の単語の生成確率．
[5] 未知語は解析する場合サ変名詞として確率付与される．
[6] 訂正実験は学習した OCR テキストに対して行なうので，未知の長さの語に対する確率値の補間を行なわなかった．

2.3.2 (ロ) 文字 n-gram による未知語の抽出

上記の形態素解析システムで獲得する方法では既に辞書にある語に誤って適合するとうまく単語としての確率が獲得されない[7]．よって辞書の善し悪しに左右されない方法として文字 n-gram 頻度 [1] をOCRテキストから獲得し，その中から有効な語を抽出する．取り出した未知語は頻度を数え直して前節の形態素解析システム中の未知語として同様に登録する．

まず，抽出したい単語の条件を示す．

- 2 字以上である．
- ひらがなを含まない．
- 辞書に登録されていない．

これに適合する文字列を n-gram の中から取り出す．手順を以下に示す．

1. OCR テキストから長さ $n = 2$ 以上の文字 n-gram 頻度を計算する．
2. 低頻度[8] の n-gram を捨てて最長の文字列[9] を取り出す．
3. 2. からひらがなを含まない文字列を取り出し，字種ごとに分けた単語も未知語とみなす (図 2 参照)．
4. 品詞 trigram モデルの辞書を調べ重なる語は排除する．

次に，未知語のコーパス中の頻度を数える．未知語を記録しておきOCRテキストを先頭から順に調べて出現回数を数える．ただし，2 重に頻度を数えないように文中の文字列で未知語として数えた部分では他の語は数えない．また，必ず最長に一致する未知語で頻度を数える[10]．

このようにして取り出された未知語を (9) 式の $C(w,$ サ変名詞) として代入し選択の言語モデルとした．このとき方法 (イ) で獲得した未知語と入れ換えて使用する．

[7] 例えば複数の漢字で構成された単語などが単漢字に分割される場合など．
[8] 頻度 4 以下を捨てた．これも獲得された n-gram に誤り文字列が多く含まれているため行った．
[9] ある文字列が抽出されたとき，その文字列が他の部分文字列になってない文字列のこと．
[10] 未知語として “ウイルス” と “ウイルス遺伝子” があり，文中で「ウイルス遺伝子属性が..」とあった場合 “ウイルス遺伝子” を 1 つと数え「属性が..」の部分で最長の未知語を調べる．なければ 1 文字づつスキップする．これを繰り返す．

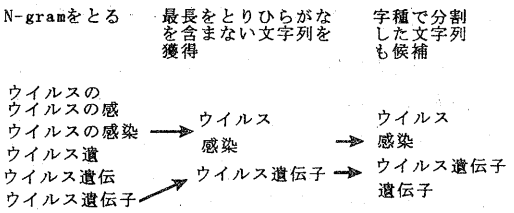

図 2: 文字 n-gram 統計からの未知語の抽出

3 OCRテキストに対する訂正実験

大量なテキストコーパスとして日経新聞９４年から３ヶ月分の記事（約４０万文）を用いた．また，実際の OCR 処理されたテキストデータとして奈良先端科学技術大学院大学電子化図書館で蓄積しているバイオサイエンス関連の論文誌に対するＯＣＲ読み取り結果 (約 7000 文) を用いた．

誤り訂正能力を評価するためのテストテキストは学習に用いたＯＣＲテキストの中から３ページを選んだ．絵の部分や，段組獲得の失敗により大きく誤っている文は除いた．テキストには挿入，置換，削除誤りが存在する．そこで正解テキストを人手により作成し，適合率，再現率と置換誤り改善率により評価を行う．適合率と再現率の式は以下の通りである．

$$\text{適合率} = \frac{\text{訂正後の正解の文字数}}{\text{ＯＣＲテキストの文字数}} \times 100$$

$$\text{再現率} = \frac{\text{訂正後の正解の文字数}}{\text{正解テキストの文字数}} \times 100$$

置換誤り改善率は訂正の結果，置換誤りの総数がどれだけ減ったかを表す．(誤り文字が正しい文字に置換された数) を (e) とし（正しい文字が誤り文字に置換された数）を (f) としると

$$\text{置換誤り改善率} = \frac{(\text{e}) - (\text{f})}{(\text{入力テキストの置換誤り数})} \times 100$$

となる．

テストテキストの特性を表１に示す[11]．以下ではこれら３つのテストテキストを１つにまとめたテキスト g について誤り訂正の報告を行う．

[11] テキスト d, e, f は論文誌の１ページを表しており，内容は大腸菌の細胞，毒素に関連したものが記述されている．

表 1: テストテキストの特性

	テストテキスト			
	d	e	f	g(total)
総文字数	1641	1519	1126	4286
置換誤り	62	70	45	177
挿入誤り	3	2	0	5
削除誤り	3	5	8	16
正解文字数	1576	1447	1081	4104
再現率	96.0	95.1	95.2	95.5
適合率	96.0	95.3	96.0	95.6

我々の構築したシステムは誤りを含むデータから学習するので，訂正によって不用意な改悪がしばしば起こる．そこで，以下のような制約を入れた．

- 英字，数字に関して訂正を行わない[12]．
- ひらがな文字の候補は選択しない[13]．

訂正実験は 候補を出す際，２文字連続まで行う場合と１文字のみ出力する場合 (文献 [11] 参照) を行った．また，(7) 式の単語変換確率について $\alpha = \beta =$ 0.0001, 0.0005, 0.001 の３つの場合を調べた．以下，未知語の獲得方法 (イ) と (ロ) の場合の訂正実験結果について表 2，表 3 と表 4 に示す．

表 2: ＯＣＲ誤り訂正結果 ($\alpha = \beta = 0.0001$)

候補の連続文字	単語の獲得方法	テストテキスト g		
		再現率	適合率	置換率
1	(イ)	95.9	96.1	10.2(18/176)
	(ロ)	95.8	96.1	8.0(14/176)
2	(イ)	95.6	95.9	2.8(7/176)
	(ロ)	95.8	96.0	6.2(11/176)

表 3: ＯＣＲ誤り訂正結果 ($\alpha = \beta = 0.0005$)

候補の連続文字	単語の獲得方法	テストテキスト g		
		再現率	適合率	置換率
1	(イ)	95.6	95.9	2.8(5/176)
	(ロ)	95.7	96.0	5.7(10/176)
2	(イ)	95.6	95.8	1.7(3/176)
	(ロ)	95.7	96.0	4.5(8/176)

上記表中の置換率は置換誤り改善率を示しており，括弧内の分子の数字は訂正された文字の総数である．候補文字の再現率は表には示していないが，

[12] 英字と数字はＯＣＲ処理テキストでかなり誤りが多く，辞書の記述にもあまり存在しないので正確な訂正が行えないためである．

[13] ひらがなは高確率の助詞などに改悪されることが予備実験の段階で多かったためである．

表 4: ＯＣＲ誤り訂正結果 ($\alpha = \beta = 0.001$)

候補の連続文字	単語の獲得方法	テストテキスト g		
		再現率	適合率	置換率
1	(イ)	95.6	95.8	1.7(3/176)
	(ロ)	95.7	96.0	5.1(9/176)
2	(イ)	95.5	95.8	0.57(1/176)
	(ロ)	95.6	95.9	2.8(5/176)

テキスト gで1文字候補のとき 97.8%，2文字候補のとき 98.0% でであった．

誤り訂正の結果から，表2の場合に (イ) と (ロ) のどちらの精度とも最も良かった．特に (イ) の方法の結果が大きく良い．置換誤り改善率では1文字候補の時 10.2%の改善を示した．しかし，表2の下段の2文字候補の場合半分以下の精度となる[14]．またα と βの値を大きくなるにつれて (表2から表4に向かうにつれて) 急激に置換率が低下する．誤り箇所の原因は辞書に未登録な2漢字が他の既存の単語に変換されるなどがあった[15]．これに対して (ロ) のモデルは2文字候補の場合やα と βの値を大きくしても急激に悪くなることがなく比較的安定している．各テキスト d,e,f についてもほぼ均等に訂正能力を示した．この結果から最高の改善精度は (イ) の方が優れていたが手法 (ロ) の方が安定していて手法 (イ) よりも未知語が獲得できていることがわかる．

手法 (ロ) の置換誤り改善率は表2の場合2文字候補の場合で 6.2%である．文献 [6] における新聞記事を用いた疑似誤り訂正実験の結果おいて 95%のテキストの時，品詞 trigram モデルで置換誤り改善率は 27.2% であった．このことから，同じ分野の大量テキストデータが存在しない場合，システムの訂正能力が約 1/4 程度になったことがわかる．

この実験は学習したテキストデータを実験に使用しているインサイドデーダで行っている．この手法は誤りを含むテキストデータからの学習を行っているのでこの評価を用いたが，アウトサイドで行った場合には未知語が増加するためにほとんど改善されない．アウトサイドデータで評価できる頑健なシステムにするには他の情報（専門分野の単語辞書など）が必要となるであろう．

[14] 表には示していないが各テキストに対する精度が大きく異なりテキスト d に対して大きく改悪しているのが原因である．

[15] 例えば「産生」という単語が辞書に無く，また学習の際「産」と「生」の単漢字として学習されたため，未知語扱いとなり，既知の「発生」などに置き換わった．

4 考察

Tong[7] らは言語モデルと OCR モデルを同時に再推定する方法を提案している．OCR モデルを再推定するために英語文字間の誤り特性を画像的な情報を使用せずに再推定している．しかし，この手法は日本語文字種の多さ (約 3000 種) や2文字単語が非常に多い事などから日本語には適していない．また，彼らは挿入や削除誤りを実現しているが，単語内のみの比較であるため，日本語のように単語間が空白無く接着する場合の誤りは修正できなかったと報告している．この点において我々は境界が異なる単語候補を動的に扱っているので彼らの方法より優れているといえる．

日本語の OCR 誤り訂正において統計的な言語モデルと文字の画像的な類似度を用いる永田 [9] の手法がある．永田の手法では言語モデルの学習には EDR コーパスを用いて，他分野に対して訂正実験を行ない，訂正能力を示している．しかしながら分野の異なりによって生じる辞書に無い未知語に対する獲得は行なわれていない．我々の手法では他分野における，言語モデルの学習法，ならびに辞書の獲得方法について議論し，実験結果によりその有効性を示した．この点において我々のモデルの応用性が高い．

5 まとめ

解析対象と同分野のテキストデータがない場合，分野の一致しない大量テキストデータと解析対象のＯＣＲ処理されたテキストを利用する訂正システムを構築した．大量テキストデータから，文字 trigram と品詞 trigram モデルの品詞接続確率を推定し，OCR処理されたテキストから，文字 trigram，未知語，単語の生成確率を獲得する手法を提案し，訂正実験を行った．未知語の獲得方法として形態素解析システムのみを利用する場合と文字 n-gram 統計量も利用する場合について実験した．実際のＯＣＲ処理されたバイオサイエンス関係の論文誌のテストデータに対して，テキストを改善することが確認された．その結果，置換誤り改善率において，解析対象と同分野の大量コーパスで学習した場合の約 1/4 の訂正能力を示した．

解析対象と同分野のテキストデータがない場合，

今回は，テキストデータからの学習だけでどれだけ誤り文字列を訂正できるかを示した．今後は文字の挿入，削除誤りを考慮したい．また，言語モデルとして距離が離れた単語間の影響を考慮したモデルが提案されており [2][5] モデル化において参考にしたい．

6 謝辞

新聞記事を使用させていただいた日経新聞社，ならびにＯＣＲ処理後のテキストデータを使用させて頂いた奈良先端科学技術大学院大学電子化図書館に感謝の意を表します．

参考文献

[1] Gonnet, G. H., Baeza-Yates, R. A. and Snider, T.: New Indices for Text: PAT Trees and PAT Arrays, *Informatioon Retrival: Data Structures and Algorithms*, pp. 66–82 (1992).

[2] Kuhn, R. and Mori, R. D.: A Cache-Based Natural Language Model for Speech Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, No.6*, pp. 570–583 (1990).

[3] Kukich, K.: Techniques for Automatically Correcting Words in Text, *ACM Computing Surveys 24*, pp. 377–439 (1992).

[4] Nagata, M.: Context-Based Spelling Correction for Japanese OCR, *Proc. COLING-96*, pp. 806–811 (1996).

[5] Rosenfeld, R.: A Maximum Entropy Approach to Adaptive Statisical Language Modelling, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.10, No.4*, pp. 187–228 (1996).

[6] Takeuchi, K. and Matsumoto, Y.: Japanese OCR Error Correction Using Stochastic Morphological Analyzer and Probabilistic Word N-gram Model, *Proc. ICCPOL'99* (1997). To appear.

[7] Tong, X. and Evans, A. D.: A Stitistical Approach Automatic OCR Error Correction in Context, *Proc. 4th Very Large Corpora*, pp. 88–100 (1996).

[8] Webster, R., 中川正樹: 英語と日本語を対象にした文章誤り検出・共通点と 相違, 情報処理, *vol.37, No.9*, pp. 865–871 (1997).

[9] 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, 電子情報通信学会論文誌, *VOL.J81-D-II,No.11*, pp. 2624–2634 (1998).

[10] 竹内孔一, 松本裕治: 隠れマルコフモデルによる日本語形態素解析の パラメータ推定, 情報処理学会論文誌, Vol. 第 38 巻, No. 第 3 号, pp. 500–509 (1997).

[11] 竹内孔一, 松本裕治: 共起情報と統計的形態素解析による OCR 誤り訂正, 情報処理学会 自然言語処理研究会 *NL121-3*, pp. 17–24 (1997).

[12] 久光徹, 丸川勝美, 嶋好博, 藤澤浩道, 新田義彦: OCR 誤認識後処理の効率について, 情報処理学会自然言語処理研究会 *104*, pp. 17–24 (1994).

[13] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明: 日本語形態素解析システム『茶筌』version 1.0 使用説明書, NAIST Technical Report NAIST-IS-TR97007 (1997).