統計的言語モデルを用いたOCR誤り訂正システムの構築

竹内孔一[†] 松本裕治[††]

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．日本語におけるOCR誤り訂正の先行研究には，OCRの文字候補と品詞タグ付きコーパスを利用した研究がある．しかしながら，分野が異なれば単語の出現分布などが変わることから，誤り訂正を行う分野と同分野のタグ付きコーパスを用意する必要があり，それには大変コストがかかる．また，分野によっては統計学習に必要な電子化テキストデータがない場合も多い．そこで，まず我々は学習用として電子化された大量テキストデータを仮定したOCR誤り訂正システムを構築し，ランダムに生成された文字置換誤りテキストに対する訂正実験を行った．次に，電子化テキストがない分野に対して，OCR処理された誤りを含むテキストを学習に利用するシステムを作成し評価を行った．システムは，文字trigram，統計的形態素解析システム，単語trigramを用いた．大量テキストを仮定したシステムでは，90%の文字読み取り精度のテキストを92.9%まで改善し，95%の精度のテキストを96.4%にまで改善した．また，電子化テキストデータがない場合について，実際のOCR処理されたテキストに対する訂正実験を行い，その有効性を示す．

OCR Error Correction Using Stochastic Language Models

KOICHI TAKEUCHI[†] and YUJI MATSUMOTO[††]

In recent years, OCR error correction is getting more and more important for the purpose of converting printed texts into electronic ones on computers. As a previous work, there exists a study of OCR post processing which uses OCR's character candidates and a morphological analyzer trained on part-of-speech-tagged corpus. However, too much cost is required to prepare pos-tagged corpus for each domain. In this paper, we present an OCR error correction method which uses stochastic language models trained on large texts. We also construct an OCR error correction system which uses OCR's output texts in a domain in which no large scale training text exists. Our system consists of the models of character trigram, a stochastic morphological analyzer and word trigram. We show that the models trained on large texts improve a text of 90% correct character rate into that of 92.9% correct rate and a 95% correct text into a 96.4% correct one. We also show how the models trained on OCR's output texts correct errors in the OCR's output texts.

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が世界中の人々から利用できるようになってきた．その中で文字情報は基本的な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（Optimal Character Reader，光学式文字読み取り装置）を利用することが多い．OCRは文字に関する画像情報をもとに文字画像を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．本論文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することを目的とする．

本論文では，統計的手法を用いて，テキストデータから学習した言語モデルを利用した誤り訂正システムを提案し，誤りを含むテキストに対する訂正実験の結果を報告する．この際，文字間の画像的な類似度は利用しなかったため，本手法はOCR後処理だけでなく誤字訂正にも応用できる．

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[4]にまとめられている．英語は日本語と異なり，単語間が空白によって区切られているため，単語単位で誤り訂正を行うことができる利点がある[9]．Tongら[8]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

† 学術情報センター
National Center for Science and Information Systems
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

確率モデルで候補を選択するモデルを提案し，効果をあげている．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単位ではなく文字単位のn-gramを応用した確率モデルが日本語文の文字誤りの検出・訂正に対して有効であることが示されている[10),12),15)]．しかし，文字n-gramでは文法的な訂正情報を反映することができない．また，文字種が英語に比べて多いことから，漢字やカタカナなど文字種に応じた訂正法に関する研究も行われている[11),17)]が，これらの方法は文章中から単語を取り出した後の処理手法であるため，文中の単語の境界の曖昧性の問題を解決していない．これに対して，文法的な制約として形態素解析モデルを利用し，複数の単語列候補の中から動的に最適な単語列を選択する手法[5),16)]が提案されている．

特に，Nagata[5)]は，OCRの文字候補付きの出力と品詞trigramに基づく形態素解析システムを用いて，未知語を含んだテキストデータに対して誤り訂正実験を行い高い精度を得ている．この方法では統計的形態素解析システムを学習するために解析対象となる文章と同分野のタグ付きコーパスが必要となる．しかしながら，誤り訂正を行う分野と同分野のタグ付きコーパスを用意するには大変コストがかかる☆．また，分野によっては電子化されたテキストデータがない場合も多い．そこで，本論文では以下の2つの条件について，誤り訂正システムの構築と誤り訂正実験を行った．

1 誤り訂正するテキストと同分野の大量テキストデータの存在を仮定する．
2 同分野の大量テキストデータは存在せず，OCR処理されたテキストのみ存在すると仮定する．

本論文で提案する誤り訂正システムは基本的に以下の3つの部分から構成されている．最初に，1）文字誤り箇所検出を行い，次に，2）文字候補を生成して辞書引きにより単語候補の作成を行う．最後に，3）単語列候補の選択を行う．文字誤り検出と文字候補の生成には文字trigramを利用し，候補の選択には，品詞trigramモデルと単語trigramモデルを利用した．ただし，文字の置換誤りのみを対象とし，挿入と削除の誤りには対応していない．また，文字間の画像的な類似度は利用していない．2つの条件下では，これらの各部分で用いられる統計モデルと学習法が異なる．

誤り訂正の対象と同分野の大量テキストデータを仮定した場合（**1**の場合）は新聞記事データを題材にした．文字誤り検出には文字trigramモデルを使用し，候補選択に関しては上記のすべての統計モデルを用いてシステムを構築した．誤り訂正実験は，学習に使用しなかった新聞記事に乱数を用いて，置換誤りを生成したデータに対して行った．訂正実験の結果，90%の文字読み取り精度のテキストを92.9%まで改善し，95%の精度のテキストを96.4%に改善した．

誤り訂正の対象と同分野のテキストデータがない場合（**2**の場合）には，文字の連接確率や単語の頻度分布が正確に推定できない．そこで，他分野であるが，大量テキストで獲得された統計量（**1**の場合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の統計量を融合してモデルを構築する．誤りを含むテキストから辞書に未登録な単語を獲得し，出現頻度を推定する手法も提案する．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ことを示し，**1**の場合に比べて文字の置換誤りの改善精度が約1/2程度となったことを報告する．

以下，2章で大量テキストデータを用いた誤り訂正システムの構築を行い，3章で乱数によって置換誤りを生成したデータに対する訂正実験の結果について述べる．4章でOCR読み取りデータからの未登録単語の獲得方法と訂正システムについて説明する．5章では実際のOCR処理されたデータに対する訂正実験の結果を示す．6章では考察を行い，7章でまとめる．

2. 大量テキストデータを用いたOCR誤り訂正システム

OCRの誤り訂正を次のように定式化する．OCRの入力文字列を $C = c_1, c_2, \ldots, c_m$ とし，その出力文字列を $X = x_1, x_2, \ldots, x_k$ 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．

$$\hat{C} = \underset{C}{arg\,max}\, P(C|X) \tag{1}$$

ベイズの定理から

$$\hat{C} = \underset{C}{arg\,max}\, \frac{P(X|C)P(C)}{P(X)} = \underset{C}{arg\,max}\, P(X|C)P(C) \tag{2}$$

となる．$P(X|C)$ はOCR装置の確率モデルである．我々の提案する手法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．

☆ 学習コーパスの分野の異なりによる影響について文献7), 13)に実験結果がある．

訂正手法は以下の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 単語列候補の選択（Selection）．

2で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．式(2)を書き直して

$$\hat{W} = \underset{W}{arg\,max}\, P(S|W)P(W) \tag{3}$$

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，候補の選択に役立つ．$P(S|W)$ はこれら単語列の混同確率を示しており以下の式で推定する．

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

n は文中の単語数を示す．w_i は候補の単語，s_i は w_i に対応するOCRの出力文字列であり，$P(s|w)$ が単語混同確率である．この式は単語候補生成の部分（2.2.3項）で計算される．以下，各モデルについての詳細を記述する．

2.1 文字誤り箇所の検出（Detection）

文字trigram確率を用いて文字誤り検出を行う方法は，文献10)，15)にあるが，本システムでは文献10)に近い方法をとる．つまり，文字列 c において確率 $P(c_i|c_{i-2}, c_{i-1})$ が Tp（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，これを文字列の先頭から順に当てはめてゆき，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（**図1**）．つまり，誤りの原因となりうる多くの文字を誤り訂正の候補と考えることになる．3章において，単純に文字trigram確率の c_i のみを誤りとした場合と比較実験を行い，再現率，適合率ともに我々の方法が有効であることを示す．

図1は文字誤り箇所検出の具体的な出力例を示している．図1下段の出力例において，×印は合計得点が -3 点の文字，△は -2 点の文字を表しており，-1 以上は○印で示している．文献14)の実験結果から×と△印まで誤り文字と判断し，文字候補の生成を行う．

2.2 単語候補の生成（Generation）

英語では単語ごとにわかち書きされているので，誤り文字を修正する際，誤った単語の文字列としての類似度により，正しい単語を類推することができる．この方法は日本語でもカタカナ文字列などに応用できる[17]が，漢字の数が大量であること，ならびに単語の境界が明示されないことから，一般的には英語の場合と同様の方法をとることはできない．そこで，単語単位ではなく，文字trigramモデルを利用して文字候補を生成する．生成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混同確率を計算する．

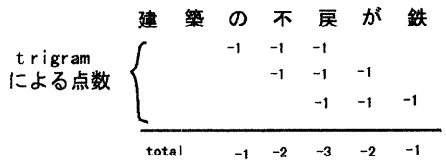

図1 文字誤り箇所の検出の出力例
Fig. 1 Example of OCR error detection.

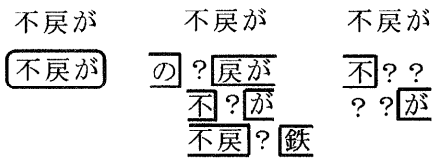

図2 誤り指摘箇所からの文字候補の生成
Fig. 2 Generation of candidates for error correction from OCR error detection.

以下では，誤り指摘箇所から文字候補を生成し，辞書引きを行い単語候補を得るまでを順に説明する．

2.2.1 誤り指摘箇所からの文字候補の生成

図1の文字誤り検出の例でもあるように，1文字（「戻」→「振」,「雨」→「業」）しか誤っていない場合でも，その周辺の文字も誤り箇所と指摘される．指摘された箇所を連続文字列誤りとして候補を推定しても，3文字以上の連続誤りとなると，ほぼ想起は不可能といえる．文字trigram確率を用いた推定では，文献14)の実験から2文字連続想起を超えると大きく精度が悪くなることが示されている．そこで，連続した誤り指摘箇所（図1では「不戻が」「工事雨者」）について，誤っている連続文字列を0から2文字列までとし，その他の文字は正しいと仮定して，すべての組合せについて以下に述べる文字列候補の生成を行う．

図2では「不戻が」の例を示している．図中の「?」の部分が誤りと仮定している部分である．これを最大2文字連続まで，文字誤り箇所検出で得られたすべての場所について行い，各部分での候補文字を生成する．

2.2.2 文字列候補の生成

文字trigramモデルを用いて文字候補の生成を行う．上述したように，最大で2文字連続までの候補を出力

する．以下は 2 文字の候補文字列の例である．候補文字列を m_i, m_{i+1} とし，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，$P(m_i|\, c_{i-2}, c_{i-1}) > 0$ を満たす候補文字 m_i を生成し，次に，各候補文字 m_i に対して $P(m_{i+1}|\, c_{i-1}, m_i) > 0$ を満たす候補文字 m_{i+1} を生成する．m_{i+1} は多数の候補が生成されるため，$P(m_i|\, c_{i-2}, c_{i-1})\, P(m_{i+1}|\, c_{i-1}, m_i)$ の確率値の上位 300 個の候補 m_i, m_{i+1} のみを残す．最後に，各候補について $P(c_{i+2}|\, m_i, m_{i+1})\, P(c_{i+3}|\, m_{i+1}, c_{i+2})$ の確率値を計算し，確率の積算値 P_f の上位 5 個の文字列 m_i, m_{i+1} を候補とする．

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

同様の手法で，文末方向から文字 trigram を計算して上位 5 個の候補を計算し，各場所について合計 10 候補作成する[14]．

2.2.3 辞書引きと単語混同確率の計算

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の辞書を使用する．辞書には約 10 万語の単語が登録されている．**図 3** は辞書引きの例を示しており，最上位のパスは最も確率値の高かったパスを示している．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混同確率を計算する．単語候補 w_i に対して，$L(w_i)$ 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 OCR 出力の文字列（単語）として，単語の混同確率 $P(s_i|w_i)$ を以下の式で近似した．

$$P(s_i|w_i) \approx \prod_{l \in L(w_i)} \alpha^{k_l - 1} \beta^h \quad (6)$$

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 の信頼度を表し，β^h は入力文字列と単語候補との編集距離（edit distance）を表している．

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 3 章の実験では 0.05，5 章の実験では 0.0001，0.0005，0.001 の 3 種類の値を用いた．

2.3 単語列候補の選択（Selection）

生成された単語候補から正しい単語列を選択するために，統計的形態素解析システムと単語 trigram モデルを適用する．これらの統計量の学習法について説明した後，実験に用いた言語モデルについて記述する．

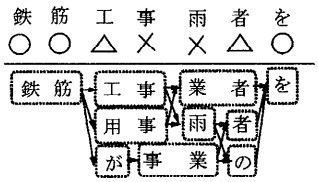

図 3 候補文字列からの辞書引き

Fig. 3 Consulting a morphological lexicon with candidate characters.

2.3.1 大量のテキストデータで学習する方法

形態素解析システムや単語の統計量を持ち込む際，問題となるのは単語のわかち書きの精度である．形態素解析は 100%の精度ではないため，誤りを含んだ単語列を生成することになる．しかし，OCR の文字誤り訂正の目的のためには，正しい単語列を生成する形態素解析が必ずしも必要というわけではない．なぜなら，同じ入力文字列に対しては確実に同じ単語列（と品詞列）を生成するため，その誤りを含んだ単語列の統計的分布を学習すれば元の文字列の特性を反映した結果を学習できることになるからである．そこで，上記の単語に関する統計モデルを学習するときには，同じ形態素解析システムでテキストデータを解析した出力結果を利用する．

形態素解析システムとしては 96%以上の精度を持つ茶筌[18]があるが，内部のコスト値が人手で作成されているため，他の統計量と同様に確率的に扱えない．そこで，テキストデータを茶筌で解析させ，出力される形態素列から品詞の連接確率と，単語の生成確率を獲得して統計的形態素解析システムを作成する．

学習は次の順序で行う．

1） 大量テキストデータを形態素解析システムで解析し，その出力結果を学習して統計的形態素解析システム（品詞 trigram モデル）を作成する．

2） この統計的形態素解析システムで同じテキストデータを解析してタグ付きコーパスを作成する．

3） タグ付きコーパスから単語 trigram を学習する．

これら獲得された統計モデルを利用して，単語列候補の選択モデルを構築する．

2.3.2 単語列候補選択の言語モデル

式 (3) 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．本項では単語列候補選択のための言語モデル $P(W)$ として以下の 3 種類のモデルを考える．

- **統計的形態素解析システム（品詞 trigram モデル）**

$$P(W) \approx P(W,T)$$

$$= \prod_{i=1}^{n+2} P(w_i|t_i)P(t_i|t_{i-2}, t_{i-1}) \quad (7)$$

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，t は品詞を表している．t_{-1}, t_0 は文頭，t_{n+1}, t_{n+2} は文末，w_{n+1}, w_{n+2} は空語を示している．このモデルは Nagata[5]が使用しているモデルである．

- **単語 trigram モデル**

言語モデルとして単語 trigram モデルを適用する．

$$P(W) \approx \prod_{i=1}^{n+2} P(w_i|w_{i-2}, w_{i-1}) \quad (8)$$

ここで w_0 と w_{-1} は文頭を表し，w_{n+1} と w_{n+2} は文末を表す．単語 trigram 確率 $P(w_i|\ w_{i-2}, w_{i-1})$ を推定するために CMU tool kit[6]を利用し，バックオフ法を用いた．

$$\begin{aligned} P(w_i|w_{i-2}, w_{i-1}) = {} & \lambda_a P(w_i|w_{i-2}, w_{i-1}) \\ & + \lambda_b P(w_i|w_{i-1}) \\ & + \lambda_c P(w_i) \end{aligned} \quad (9)$$

ここで λ_a, λ_b, λ_c は $\lambda_a + \lambda_b + \lambda_c = 1$ を満たす．このパラメータは削除補間法によって計算した．式 (9) の学習に用いるタグ付きコーパスには，統計的形態素解析システムの辞書に未登録な単語もわかち書きされて入っている．よって辞書に未登録な単語も含めた単語 trigram 確率を推定できる．

- **単語と品詞の trigram モデル**

さらに上記の品詞 trigram モデルと単語 trigram モデルを融合したモデルを考える[2]．

$$\begin{aligned} P(W) &\approx P(W, T) \\ &= \prod_{i=1}^{n+2} P(w_i|w_{i-2}, w_{i-1}, t_i) \\ &\quad \cdot P(t_i|t_{i-2}, t_{i-1}) \end{aligned} \quad (10)$$

ここで，$P(w_i|w_{i-2}, w_{i-1}, t_i)$ に対して以下のようにスムージングを行う．

$$\begin{aligned} &P(w_i|w_{i-2}, w_{i-1}, t_i) \\ &= (1 - \lambda_t)P(w_i|t_i) + \lambda_t P(w_i|w_{i-2}, w_{i-1}) \end{aligned} \quad (11)$$

単語 trigram 確率 $P(w_i|\ w_{i-2},\ w_{i-1})$ は同様に CMU tool kit を利用し，かつバックオフ法を用いて求めた．λ_t は削除補間法で値を求めた．

3. ランダム誤りテキストに対する訂正実験

学習に用いるテキストデータは日経新聞94年の記事を用いた．文字 trigram，品詞 trigram，単語 trigram の学習には3カ月分の記事（約40万文）を用いた．

訂正システムの評価のために乱数を用いて誤り文字列を生成したテストデータを2種類作成した．それぞれ，5%と10%の文字誤りを含むデータである．これに対して誤り箇所検出の評価，文字候補の評価，そして最終的なシステム全体の訂正能力を測定する．

表1 テストデータの特性
Table 1 Statistics of test data of Nikkei news paper texts.

		テストデータ
総文字数		19809
10%	誤り文字数	1973
	文字認識率	90.0
5%	誤り文字数	989
	文字認識率	95.0

3.1 誤り箇所検出の評価

テストデータとして学習に用いなかったテキスト中から1文ずつランダムに300文取り出したものを用意した．このテキストに対して乱数を用いて5%と10%の置換誤りを作成した．実際の論文誌に対するOCR処理したテキストデータを観測すると，ほぼ91～93%程度☆の精度であることから現実のOCR読み取り後のテキストに近いデータといえる．実験に使用したテストデータの特性を**表1**に示す．

誤り箇所検出のモデルとして単純な方法と我々の提案する方法を比較する．単純な方法とは文字 trigram 確率モデル $P(c_i|\ c_{i-2}, c_{i-1})$ が Tp（足切り値）以下の場合，c_i のみを誤りとする方法である．ここで $Tp = 0$ として実験を行った（2.1 節参照）．

結果を**表2**と**表3**に示す．再現率と適合率は以下に示す式で計算した．

$$\text{検出の再現率} = \frac{\text{正しく指摘された誤り文字数}}{\text{真に誤っている文字数}} \times 100$$

$$\text{検出の適合率} = \frac{\text{正しく指摘された誤り文字数}}{\text{モデルが誤りと指摘した文字数}} \times 100$$

これらの結果から，我々の提案している手法が単純な文字 trigram によるモデルよりも再現率，適合率において高いことが明らかになった．

3.2 文字候補生成の評価

文字候補生成の再現率を評価する．前節の我々のモデルが指摘した誤り箇所について文字候補を作成する．文字候補の再現率は以下の式で評価する．

☆ 実際の OCR 処理後の結果は置換誤りだけでなく，挿入，削除誤りも存在する．我々の手元にある処理結果において，絵の部分を文字に無理に解釈した誤りや 段組獲得の失敗による誤り部分を除くと 95%程度であった．

表2 90%精度のデータについての誤り検出結果
Table 2 Results of error detection for 90% correct data.

誤り検出モデル		テストデータ
我々の手法	再現率（%）	97.2
	適合率（%）	34.0
単純な手法	再現率（%）	86.6
	適合率（%）	29.6

表3 95%精度のデータについての誤り検出結果
Table 3 Results of error detection for 95% correct data.

誤り検出モデル		テストデータ
我々の手法	再現率（%）	97.7
	適合率（%）	28.9
単純な手法	再現率（%）	84.8
	適合率（%）	22.7

表4 文字候補生成の精度
Table 4 Total performance of generating candidates.

テストデータ		文字候補生成の精度
90%	再現率（%）	97.4
95%	再現率（%）	99.1

$$\text{文字候補の再現率} = \frac{\text{候補文字を含む入力文字列のうち正しい文字数}}{\text{正解テキストの文字数}} \times 100$$

実験の結果を**表4**に示す．これから90%精度のデータに対しては97%以上，95%データに対して99%の再現率で文字候補を生成することが明らかになった．

3.3 誤り訂正モデル全体の評価

評価実験において誤り検出モデルと候補生成モデルは前出のモデルで変更しないが，単語列の選択モデルとして適用する言語モデルには，(P) 品詞 trigram モデル，(W) 単語の trigram モデル，(PW) 品詞と単語の trigram モデルの3つのモデルを用意した☆．

評価は改善率，改悪率，置換誤り改善率，誤り訂正後の文字認識率で観測する．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している．各々の式を以下に示す．

(**A**) 改善率
$$= \frac{\text{誤り文字が正しい文字に置換された数 (e)}}{\text{入力テキストの総誤り文字数}} \times 100$$

(**B**) 改悪率
$$= \frac{\text{正しい文字が誤り文字に置換された数 (f)}}{\text{入力テキストの正しい文字数}} \times 100$$

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

☆ ここで，式(6)における α，β の値は 0.05 を用いた．

表5 90%精度データの訂正実験
Table 5 Results of error correction for 90% correct data.

選択モデル	評価	90% correct data
P	(A)	46.6
	(B)	3.35
	(C)	16.3 (322/1973)
	(D)	91.7
W	(A)	52.4
	(B)	2.82
	(C)	26.9 (531/1973)
	(D)	92.7
PW	(A)	52.4
	(B)	2.64
	(C)	28.6 (565/1973)
	(D)	92.9

表6 95%精度データの訂正実験
Table 6 Results of error correction for 95% correct data.

選択モデル	評価	95% correct data
P	(A)	51.4
	(B)	2.00
	(C)	13.2 (131/989)
	(D)	95.7
W	(A)	57.5
	(B)	1.88
	(C)	21.8 (216/989)
	(D)	96.1
PW	(A)	58.1
	(B)	1.64
	(C)	27.0 (267/989)
	(D)	96.4

(**D**) 誤り訂正後の文字認識率
$$= \frac{\text{訂正後の正しい文字数}}{\text{入力テキストの総文字数}} \times 100$$

表5と**表6**に90%と95%精度データの訂正実験結果を各々示す．表中の数字はすべて百分率表示をしており，評価（C）（置換誤り改善率）の括弧の中は具体的な数値を示している．

表5と表6の誤り訂正実験の結果から，(D) 誤り訂正後の文字認識率において，平均的に（PW）品詞と単語の trigram モデルが最も精度が良い．(PW) は他のモデルに対して改善率が高い割に改悪率が低くバランスがとれていることが分かる．文献5)では（P）品詞 trigram モデルを利用しているが，上記の結果から（PW）の方が効果的である．(C) 置換誤り率で評価すると90%精度データに対して（表5）（PW）では28.6%で，95%精度データ（表6）の（PW）では

27.0%となっており，データ中の誤り率の増減に対してほぼ変わらない精度で文字訂正が行われていることが分かる．訂正実験の結果として（PW）の場合で90%，95%精度のデータがそれぞれ92.9%，96.4%精度のデータに訂正された．

4. 他分野のデータに対するOCR誤り訂正システムの構築

OCR処理されたテキストデータ（以降，OCRテキスト）を訂正する場合，大量の電子化された同じ分野のテキストデータが存在する場合は少ないといえる．そこで，大量テキストで学習した統計量と，誤りを含むOCRテキストから獲得した統計量を利用することにより，他分野のデータに対するOCR誤り訂正システムを構成する手法を提案し，実験結果を示す．

提案する訂正手法は前章と同様に，1）誤り箇所検出，2）単語候補の生成，3）単語列候補の選択からなる．OCRテキストからの学習は，2）における文字候補生成のtrigram確率と，3）における未知語を含めた単語の出現確率の獲得について行った．

以下の節で，それぞれの処理についてどの統計量を使用するか説明する．

4.1 誤り箇所検出

誤り箇所検出の文字trigramモデルの統計量の獲得は，OCRテキストの分野と異なる大量テキストから行う．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ことになる．

4.2 文字候補の生成

文字候補を生成する文字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かった．

2 未出現文字列に対して最低確率値 P_{min} を設けた．

1ではしきい値（$T=4$）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（1～4）な文字列に誤りが多く含まれていたからである．**2**の P_{min} は以下のように求める．まず，未出現文字列に配分する確率の総和 P_{smc} を2文字連続の次に低頻度の文字が出現する確率として以下の式で求める．

$$P_{smc}=\frac{\sum_{C(c_i,c_{i-2},c_{i-1})\leq T}1}{\sum_{C(c_i,c_{i-2},c_{i-1})\leq T}C(c_{i-2},c_{i-1})} \quad (12)$$

ここで $C(\cdot)$ はテキスト中の文字列の頻度を表す．このとき，低頻度の文字trigramを1回出現とした．これを各低頻度の文字に分配して（$\sum_{C(c_i,c_{i-2},c_{i-1})\leq T}1$ で割って）未出現文字列の確率値 P_{min} とした．

$$P_{min}=\frac{1}{\sum_{C(c_i,c_{i-2},c_{i-1})\leq T}C(c_{i-2},c_{i-1})} \quad (13)$$

よって文字候補生成に利用する文字trigram確率は

$$P_{ocr}(c_i|c_{i-2},c_{i-1})=\begin{cases}\dfrac{C(c_i,c_{i-2},c_{i-1})}{C(c_{i-2},c_{i-1})}(1-P_{min}) & C(c_i,c_{i-2},c_{i-1})>T\\ P_{min} & C(c_i,c_{i-2},c_{i-1})\leq T\end{cases} \quad (14)$$

となる．これを2.2.2項の式(5)に代入し，同様に文字候補を生成し，辞書引きにより単語候補を生成する．

4.3 単語列候補選択の言語モデル

単語列候補選択の言語モデルには品詞trigramモデルを適用する．品詞trigramモデルは品詞の接続確率と単語の生成確率を含む．品詞の接続確率は分野による影響が少ないと考え，大量テキストから学習する．しかし，単語の生成確率は分野により異なるため，OCRテキストから学習する．このとき，分野が異なるために，辞書にない未知語を獲得し，確率値を推定する必要がある．そこで，以下の2通りの方法を提案する．

（イ） 統計的形態素解析システムを利用して未知語を抽出し，確率値を与える．

（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以下で説明する．

4.3.1 （イ）統計的形態素解析システムによる未知語を含む単語生成確率の獲得

統計的形態素解析システムは辞書にな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テキストを統計的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 $P(w|t)$☆を獲得する．統計的形態素解析システムは未知語に対する品詞の接続確率としてサ変名詞を仮定している．そこで統計的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単語の頻度を数え上げた．ただ

☆ 式(7)中の単語の生成確率．

し，すべての単語に対して，低頻度（しきい値 $Tt=4$ 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．つまり，サ変名詞以外の確率は以下のように付与される．

$$P(w|t)=\begin{cases}\dfrac{C(w,t)}{C(t)} & C(w,t)>Tt\\ P(w|\text{サ変名詞}) & C(w,t)\leq Tt\end{cases} \tag{15}$$

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

$$P_{unk}=\frac{\sum_{C(w,t)\leq Tt}1}{\sum_{C(w,t)>0}C(w,t)} \tag{16}$$

ここで $\sum_{C(w,t)\leq Tt}1$ は頻度1から Tt で出現した単語をすべて頻度1と見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．よってサ変名詞の場合は以下のようになる．

$$P(w|\text{サ変名詞})=\begin{cases}\dfrac{C(w,\text{サ変名詞})}{C(\text{サ変名詞})}(1-P_{unk}) & C(w,\text{サ変名詞})>Tt\\ \dfrac{Leng(w)}{\sum_{C(w,t)\leq Tt}1}P_{unk} & C(w,t)\leq Tt\end{cases} \tag{17}$$

上式において未知語に対して単語長の相対頻度分布関数 $Leng(w)$ をかけた．文献5)ではポア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☆1．

$$Leng(w)=\frac{\sum_{w',t;|w'|=|w|}C(w',t)}{\sum_{w,t}C(w,t)} \tag{18}$$

ここで $|w|$ は単語 w の長さを表している．

4.3.2 （ロ）文字 n-gram による未知語の抽出

上記の統計的形態素解析システムを利用して獲得する方法では，未知語がすでに辞書にある語の組合せに誤って一致すると未知語として獲得されない☆2．よって辞書の良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，前項の品詞 trigram モデルの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．

☆1 訂正実験は学習した OCR テキストに対して行うので，未知の長さの語に対する確率値の補間を行わなかった．

☆2 たと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．

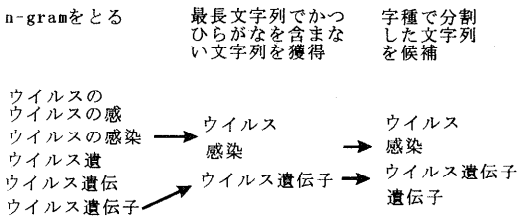

図4 文字 n-gram 統計からの未知語の抽出

Fig. 4 Extraction of unknown words using n-gram statistics.

- 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 $n=2$ 以上の文字 n-gram 頻度を計算する．

2 低頻度☆3の n-gram を捨てて最長の文字列☆4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，字種ごとに分けた文字列も未知語と見なす（図4）．

4 形態素解析用の辞書を調べ，一致する語は排除する．

次に，テキスト中の未知語の頻度を数える．未知語を記録しておき，OCR テキストを文の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☆5．

このようにして取り出された未知語とその頻度分布を式(15)から式(18)の $C(w,$サ変名詞$)$ として代入し，単語の生成確率 $P(w|t)$ を求めた．このとき，方法（イ）で獲得した未知語はまったく使用せず，n-gram を利用して求めた未知語のみを使用した．

5. 他分野のデータに対する訂正実験

3章で用いた大規模テキストデータとは異なる分野として，奈良先端科学技術大学院大学電子図書館で蓄積されたバイオサイエンス関連の論文誌に対する OCR 読み取り結果を用いた．全体で 600 Kbyte，約

☆3 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．

☆4 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．

☆5 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文字ずつスキップする．これを繰り返す．

表7 テストデータの特性
Table 7 Statistics of test data of bioscience journal texts.

	テストデータ
総文字数	4286
置換誤り	177
挿入誤り	5
削除誤り	16
正解文字数	4104
再現率	95.5
適合率	95.8

7000文である．

誤り訂正能力を調べるためのテストデータは学習に用いたOCRテキストの中から3ページ分を選んだ．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とおりである☆1．

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

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

テストデータの特性を**表7**に示す☆2．以下では，テストデータについて，誤り訂正実験の結果を報告する．

我々の構築した手法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☆3．
- ひらがな文字の候補は選択しない☆4．

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を行った．また，式(6)の単語混同確率について $\alpha = \beta = 0.0001$, 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）の訂正実験結果を**表8**，**表9**と**表10**に示す．

上記表中の置換率は置換誤り改善率のことを示しており，括弧内の分子の数字は訂正された文字の総数である．文字候補生成の再現率は表には示していないが，テストデータに対して，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．

☆1 3章の場合，適合率と再現率は同じであり，誤り訂正後の文字認識率として示している．

☆2 内容として，大腸菌の細胞，毒素に関連したものが記述されている．

☆3 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．

☆4 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8 OCR誤り訂正結果（$\alpha = \beta = 0.0001$）
Table 8 Results of OCR error correction.

候補の連続文字	単語の獲得方法	テストデータ 再現率	適合率	置換率
1	（イ）	95.9	96.1	10.2 (18/177)
	（ロ）	95.8	96.1	7.9 (14/177)
2	（イ）	95.6	95.9	4.0 (7/177)
	（ロ）	95.8	96.0	6.2 (11/177)

表9 OCR誤り訂正結果（$\alpha = \beta = 0.0005$）
Table 9 Results of OCR error correction.

候補の連続文字	単語の獲得方法	テストデータ 再現率	適合率	置換率
1	（イ）	95.6	95.9	2.8 (5/177)
	（ロ）	95.7	96.0	5.6 (10/177)
2	（イ）	95.6	95.8	1.7 (3/177)
	（ロ）	95.7	96.0	4.5 (8/177)

表10 OCR誤り訂正結果（$\alpha = \beta = 0.001$）
Table 10 Results of OCR error correction.

候補の連続文字	単語の獲得方法	テストデータ 再現率	適合率	置換率
1	（イ）	95.6	95.8	1.7 (3/177)
	（ロ）	95.7	96.0	5.1 (9/177)
2	（イ）	95.5	95.8	0.56 (1/177)
	（ロ）	95.6	95.9	2.8 (5/177)

誤り訂正の結果から，$\alpha = \beta = 0.0001$ の場合（表8）に（イ）と（ロ）のどちらの精度とも最も良かった．特に（イ）の結果が良い．置換誤り改善率では1文字候補のとき10.2%の改善を示した．しかし，表8の下段の2文字候補の場合半分以下の精度となる．また α と β の値を大きくする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文字単語が他の既存の単語に変換されるなどであった☆5．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．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分かる．

手法（ロ）の置換誤り改善率は $\alpha = \beta = 0.0001$ の場合（表8），2文字候補の条件で6.2%である．3章で，訂正対象と同分野の大量なテキストデータがある場合の5%誤りを含むテストデータに対する統計的形態素解析システムを訂正能力は，置換誤り改善率で

☆5 たとえば「産生」という単語が辞書になく，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」に置き換わった．

13.2%であった（表6）．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，本手法の訂正能力が約1/2程度になることが分かる．

この実験では，学習したテキストデータを評価に使用しているため，クローズドデータで実験を行っている．本手法は誤りを含むテキストデータから学習を行っているので，この評価を用いたが，オープンデータ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．オープン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．

6. 考　察

6.1 ランダム誤りテキストについて

前半の大量なテキストデータによる学習のOCR誤り訂正モデルの評価に対して乱数による文字置換誤りテキストを使用した．これは実際のOCRの文字誤り傾向とは一致していないが，様々な認識率のデータを簡単に作成できるので精度の評価には有効である．

乱数による文字置換は実際のOCRの文字誤り特性とは異なるが，我々は文字混同行列を用いていないので訂正精度の評価に対する影響は少ないと考えられる．ただ，低頻度な文字（ほとんど使われない単漢字など）に置換されると誤り文字検出の精度に影響があるため，ランダムに置換する文字はコーパス中に出現した文字を選択した．本研究では，文字混同行列を仮定していないが，対象とするOCRシステムの文字混同行列が入手可能なら，文字候補生成にこれを用いることができ，よりきめ細かな文字候補生成を行うことが可能であると考えられる．

6.2 他の誤り訂正モデルとの比較

Tongら[8]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こと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なく接続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誤り訂正において形態素解析システムを用いるNagata[5]の手法がある．永田の手法では言語モデルの学習には同分野のコーパスを用意する必要がある．しかし，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範囲が広い．

7. ま と め

大量テキストデータを用いた誤り訂正手法を提案した．訂正手法は，誤り箇所検出，単語候補の生成，単語列候補選択の3つの部分からなり，OCRから第1解の文字列を受け取り，訂正した文字列を出力する．使用する統計モデルは文字trigramモデル，統計的形態素解析システム，単語trigramモデルである．訂正は文字置換誤りのみを対象とし，挿入，削除には対応していない．乱数による疑似誤りテキストを訂正する実験において，高い改善能力があることを示した．この手法は，大量のテキストデータさえあれば，実装できるので分野適応が容易であると考えられる．

次に，解析対象と同分野のテキストデータがない場合，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統計的形態素解析システムを利用する場合と文字n-gram統計量を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．

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[3),7)]，今後のモデルの拡張において参考にしたい．

謝辞 新聞記事を使用させていただいた日本経済新聞社，ならびにOCR処理後のテキストデータを使用させていただいた奈良先端科学技術大学院大学電子図書館に感謝の意を表します．また，様々な貴重なコメントをいただいた査読者の方々に心から感謝します．

参 考 文 献

1) Gonnet, G.H., Baeza-Yates, R.A. and Snider, T.: New Indices for Text: PAT Trees and PAT Arrays, *Information Retrieval: Data Structures and Algorithms*, pp.66–82 (1992).
2) Jelinek, F.: Self-Organized Language Modeling for Speech Recognition, *Readings in Speech Recognition*, pp.450–506 (1990).
3) Kuhn, R. and Mori, R.D.: A Cache-Based Natural Language Model for Speech Recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.12, No.6, pp.570–583

(1990).
4) Kukich, K.: Techniques for Automatically Correcting Words in Text, *ACM Computing Surveys*, 24, pp.377–439 (1992).
5) Nagata, M.: Context-Based Spelling Correction for Japanese OCR, *Proc. COLING-96*, pp.806–811 (1996).
6) Rosenfeld, R.: The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation, *Proc. ARPA Spoken Language Technology Workshop*, pp.47–50 (1995).
7) Rosenfeld, R.: A Maximum Entropy Approach to Adaptive Statistical Language Modeling, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.10, No.4, pp.187–228 (1996).
8) Tong, X. and Evans, A.D.: A Statistical Approach Automatic OCR Error Correction in Context, *Proc. 4th Very Large Corpora*, pp.88–100 (1996).
9) 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).
10) 松山高明，渥美清隆，増山 繁：n-gramによるOCR誤り検出の能力検討のための適合率と再現率の推定に関する実験と考察，言語処理学会第2回年次大会発表論文集，pp.129–132 (1996).
11) 伊藤信泰：Bigramによるオンライン漢字認識の文脈後処理手法，情報処理学会自然言語処理研究会，NL97-6, pp.37–44 (1993).
12) 森 大毅，阿曽弘具，牧野正三：2重マルコフモデルを用いた日本語文書認識後処理，情報処理学会自然言語処理研究会，NL102-12, pp.89–96 (1994).
13) 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.38, No.3, pp.500–509 (1997).
14) 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).
15) 荒木哲郎，池原 悟，塚原信幸：2重マルコフモデルによる日本語文の誤り検出並びに訂正法，情報処理学会自然言語処理研究会，NL97-5, pp.29–35 (1993).
16) 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化について，情報処理学会自然言語処理研究会，NL104-3, pp.17–24 (1994).
17) 畑田 稔，遠藤裕英：日本語OCR文における英字・カタカナのスペル誤り訂正法，情報処理学会論文誌，Vol.38, No.7, pp.1317–1327 (1997).
18) 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).

(平成9年11月14日受付)
(平成11年3月5日採録)

竹内 孔一（正会員）

昭和43年生．平成7年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士前期課程修了．平成10年同大学院博士後期課程修了．同年より学術情報センター助手，現在に至る．統計的手法に基づく自然言語処理に関する研究に従事．電子情報通信学会会員．

松本 裕治（正会員）

昭和30年生．昭和52年京都大学工学部情報工学科卒業．昭和54年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．昭和59～60年英国インペリアルカレッジ客員研究員．昭和60～62年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，平成5年より奈良先端科学技術大学院大学教授，現在に至る．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM各会員．