言語情報にもとづく候補文字補完を用いた文字認識後処理

三部裕史† 大森健児††

言語情報を用いて手書き文字認識の誤りを訂正する研究が行われているが，これまでの手法の多くは文字認識部からの候補文字集合から正解を選び出している．しかし，これらの手法では，候補外文字，すなわち，候補文字集合の外に正解が存在する場合には，候補外文字の箇所を正しく訂正できない．さらに，正しく訂正されない誤った文字が文脈に反映されてしまうため，その前後にある正しい文字を誤って訂正するという事態を招いてしまう．本論文では，このような事態を招く候補外文字の問題を解決して精度の良い誤り訂正処理を実現する手法として，言語情報にもとづく候補文字補完を用いた誤り訂正処理手法について述べる．本手法は，候補外文字を検出し，検出された候補外文字に対して日本語に関する知識と文字の字形パターンに関する知識を用いて新たな候補文字を生成し，それを用いて誤り訂正処理を行う．すなわち，従来手法による誤り訂正処理の結果からBigramを用いて候補外文字を検出する．検出された候補外文字に対して，単語辞書の疑似的な検索と類似文字テーブルを用いて新たな候補文字を生成する．生成された新たな候補文字により文字認識部から提示された候補文字集合を補完し，再度誤り訂正処理を行う．これにより，候補外文字の影響を受けない文字認識結果の誤り訂正処理を実現する．実際の文字認識結果を用いての実験結果によれば，言語情報にもとづく候補文字補完を用いることにより，誤り訂正処理による文字認識の正解率の増分が従来手法に比べ平均50%増大した．

Post Processing for Character Recognition by Using Candidate Supplementation Based on Linguistic Information

HIROFUMI SAMBE † and KENJI OHMORI ††

This paper describes a post processing for character recognition. As the character recognition technology has not achieved a full recognition accuracy yet, incorrect results should be corrected after the character recognition. It would be carried out by scrutinizing a character string consisting of characters each of which is the result of the character recognition so that it becomes appropriate for a Japanese sentence. By using both linguistic and statistical information, our post processing system has conquered the problem of conventional systems, where if the correct candidate is not included in the candidates obtained as the character recognition result for some input pattern, the post processing system can not obtain the correct result for this input pattern and also may change the previous or following character wrong. The error correction process of our system consists of four stages: 1) candidate selection stage where for each input pattern a single candidate, which is supposed to be the most correct, is selected among the candidates obtained from the character recognition system by using the morphological analysis with the assumption that the correct candidate is included in these candidates, 2) error detection stage where for each input pattern the system checks whether it is probable the correct candidate would not be included in these candidates, 3) the candidate generation stage where, when it is probable, new candidate characters are generated by using a word dictionary and a confusion matrix, and 4) the candidate supplement stage, the new candidates are supplemented to the candidate characters, the process goes back to the candidate selection stage and selects a single character again. According to the experiment, it has been observed that the candidate supplement method by using linguistic and statistical information brings about more than 50% increase of accuracy of the error correction.

† 法政大学大学院工学研究科システム工学専攻
Systems Engineering, Graduate School of Engineering, Hosei University
†† 法政大学工学部経営工学科
Department of Industrial Engineering and Systems Engineering, Faculty of Engineering, Hosei University

1. はじめに

手書き文字の認識に関する研究が盛んに行われ，現在では，ペンコンピュータや日本語OCRといった製品が開発され，商品化されるに至っている．しかし，

ユーザが自由に筆記した手書き文字に対しては，様々な形態の文字が現れることから高い認識精度を達成することが難しい．このため，これを解決する方法の1つとして，言語情報を用いて文字認識の誤りを訂正する知識処理が提案されている[1]．

手書き文字認識を対象とした文字認識の誤り訂正処理には，文字の並びに確率モデルを当てはめる手法[2)~4)]や，単語辞書や文法規則を用いる手法[5)~7)]などがある．これらの手法は，文字認識部から提示される候補文字集合から言語情報をもとに正解文字を選び出すことに主眼がおかれている．文字認識過程は正解文字の候補を限定する大分類と，その候補の中から正解文字を選び出す細分類に分けることができるが，上記の手法は言語情報を用いた細分類のやり直しといえる．しかし，候補外文字，すなわち，正解が候補文字集合以外の文字の場合には，つまり，文字認識での大分類が失敗した場合には，大分類についてもやり直す必要がある．

従来の手法では，候補文字集合から正解文字を探しているため，候補外文字が存在する場合にはその箇所を正しく訂正できないばかりか，正しく訂正されない誤った文字が文脈に反映されるため，その前後にある正しい文字についても誤って訂正するという事態を招いてしまう．このような問題に対処する方策として，文字認識部から提示される候補文字集合を十分に大きくすることや，類似文字テーブルから候補文字を補う[5)]ことにより候補外文字の発生する確率を小さくすることが考えられる．しかし，これらの方策では候補文字の組み合わせ数が増加してしまうばかりでなく，下位の候補文字から入力文章と関係のない単語が構成されてしまう可能性が高くなる．単語辞書の類似検索，あるいは，曖昧検索と呼ばれる手法[6)]も用いられているが，語頭が候補外文字である場合には対処していない．また，類似検索の対象が文字数の多い単語に限定されるため，1単語をできる限り短い文字列で登録するように作られた一般の形態素解析用単語辞書を用いたのでは大きな成果は得られない．

本論文では，候補外文字の問題を解決し，精度の良い手書き文字認識の誤り訂正処理を実現する手法として，言語情報にもとづく候補文字補完を用いた誤り訂正処理手法について述べる．本手法では，言語情報にもとづく候補文字補完により，言語情報にもとづいて大分類のやり直しを実現する．まず，従来手法による誤り訂正処理を行い，その結果から Bigram を用いて候補外文字を検出する．検出された候補外文字に対して，日本語に関する知識として1単語ができる限り長い文字列で検索できるように構築された単語辞書[7)]と，文字の字形パターンに関する知識として類似文字テーブルを用いて新たな候補文字を生成する．そして，生成された候補文字により文字認識結果として与えられた候補文字集合を補完し，再度従来手法による誤り訂正処理を行う．このような一連の処理により，候補外文字の影響を排除した精度の良い誤り訂正処理を実現する．

2章で誤り訂正処理の概要について述べ，3章で候補外文字の検出について，4章で新たな候補文字の生成方法について述べる．5章で実現したシステムを用いた実験結果について述べ，最後に考察する．

2. 誤り訂正処理の概要

手書き文字認識結果に対する精度の良い誤り訂正処理を実現するためには，解決しなければならない3つの主な問題がある．第1に，文字列中のすべての1文字が複数の候補文字という曖昧な形で与えられる「文字の曖昧さ」の問題．第2に，日本語の文法規則では，どのような文字の並び，あるいは，単語の並びが日本語として正しくないかが必ずしも明確に示されない「日本語が持つ曖昧さ」の問題．第3に，正解が文字認識結果の候補文字集合に含まれない場合，従来手法では正しい答が得られないという，いわゆる「候補外文字」の問題である．

本論文で述べる誤り訂正処理では，次のような基本方針に従って上記の問題に対処する．

方針1． 文字認識部から与えられる候補文字の組み合わせから文法的に正しい文字列を作り上げる．

方針2． 「文法的に正しい」文字列の中から「日本語としてもっともらしい」文字列を選び出し，これを誤りの訂正結果とする．

方針3． 候補外文字には，言語情報をもとに適切な文字により候補文字集合を補完して誤りの訂正を行う．

方針1，2に従った誤り訂正処理の実現については，既に文献7)で述べた．以下ではこれを従来手法による誤り訂正処理と呼ぶ．

従来手法による誤り訂正処理では，候補外文字が存在する場合正しい訂正を行うことができない．さらには，候補外文字の前後で正しい文字が誤って訂正されてしまう可能性がある．そこで，方針3に従って言語情報にもとづく候補文字補完を行い，候補外文字の影響を排除して誤り訂正処理を行う．

候補外文字の影響をできるだけ少なくするために新

たな候補文字を候補文字集合にむやみに追加した場合には，候補文字集合が大きくなってしまう．このとき，処理量の増大だけでなく，下位の候補文字の組合せで偶然単語が構成されるという問題が増幅されてしまう．このような問題を発生させないためには，不必要な文字に対して候補文字補完を行わないこと，不必要な文字を候補文字集合に追加しないことが重要である．つまり，どの文字が候補外文字であるかを検出，あるいは，候補外文字の可能性がある文字の範囲を可能な限り絞り込むことにより候補文字補完の適用対象を限定する必要がある．また，補完のための新たな候補文字は文脈に適した文字であり，かつ，入力字形パターンと無関係でない文字である必要がある．

本誤り訂正処理手法では，まず，候補外文字の存在を考慮せずに従来手法による誤り訂正処理を行う．次に，文字の並びに関するBigramの統計をもとに訂正結果に含まれる誤りを検出し，検出された誤りをすべて候補外文字の影響であると見なす．そして，候補外文字の影響であると見なされた文字に対して，日本語に関する知識と字形パターンに関する知識を用いて新たな候補文字を生成する．日本語に関する知識としては単語辞書を，字形パターンに関する知識としては類似文字テーブルをそれぞれ用いる．生成された候補文字により文字認識結果の候補文字集合を補完して再び誤り訂正処理を行う（**図1**）．細分類のやり直しという点では，再度誤り訂正処理を行うほかに，文字認識部へ情報を還元して，文字認識部で再度細分類を行うことも考えられ，新たな候補文字の生成は認識部への情報還元を考慮した形態で行う．

候補外文字の検出については3章で，新たな候補文字の生成については4章で述べる．

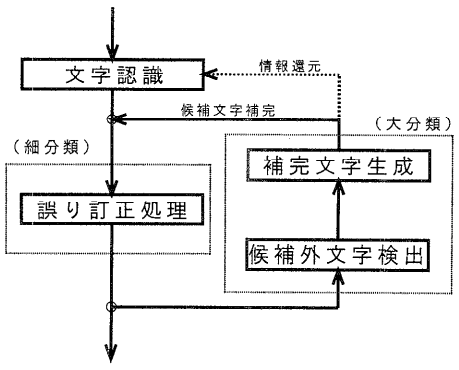


図1 言語情報にもとづく候補文字補完を用いた誤り訂正処理
Fig. 1 Schematic diagram of the error correction process by using candidates supplementation based on linguistic information.

3. Bigramにもとづく候補外文字の検出

これまでに提案されている文字認識結果の誤り訂正処理手法では，仮に文字認識結果に含まれる誤りが正しく検出できたとしても，どの文字が候補外文字であるかを識別することは行われていない．そのため，候補外文字の問題に対処する方策は，候補外文字の問題を無視するか，すべての文字が候補外文字の可能性があると考えるかのいずれかであった．そこで，本誤り訂正処理手法では，候補外文字の存在範囲を限定するための指針を与えるものとして，以下に示す2つのことを仮定する．

仮定1． 候補外文字の影響がない場合，従来手法による誤り訂正処理の適用で誤りが正しく訂正される．

仮定2． 1単語にはたかだか1文字の候補外文字しか含まれない．

2番目の仮定は，外来語や複合名詞を除く多くの単語は3文字以下の文字列であるという事実にもとづいている．候補外文字の発生確率が33%以下の場合には，1単語中に2文字以上の候補外文字が含まれる確率は十分に小さいと考えられる．また，現存する多くの文字認識システムでは，この仮定を満たすのに十分な精度が実現されていると考えられる．

第1の仮定により，従来手法による誤り訂正処理を施した結果に含まれる誤りは，すべて候補外文字か候補外文字の影響を受けていると見なすことができる．また，第2の仮定により，誤りが連続して検出された場合でも1つの単語中では1文字だけが候補外文字であり，その他は候補外文字に影響された誤りである．連続する2文字がともに候補外文字であるならば，それら2文字の間には単語の境界が存在すると見なすことができる．

そこで，文字の並びに関するBigramの統計をもとに訂正結果に含まれる誤りを検出し，新たに生成する候補文字が1単語に2文字以上含まれないように候補文字補完，および，単語抽出を行うことで適切な文字への候補文字補完を実現する．これにはまず，Bigramの統計として実際の文章において$C_i \rightarrow C_j$という文字の遷移が生じる頻度を学習する．これをもとに遷移確率$P(C_j|C_i)$を求め，Bigramを作成する．そして，従来手法による誤り訂正処理結果の文字列について，逐次文字の遷移確率$P(C_{i+1}|C_i)$を求め，閾値として適当な値Pを用いて，

$$P > P(C_{i+1}|C_i). \quad (1)$$

となるC_{i+1}およびC_iを候補外文字として検出する．

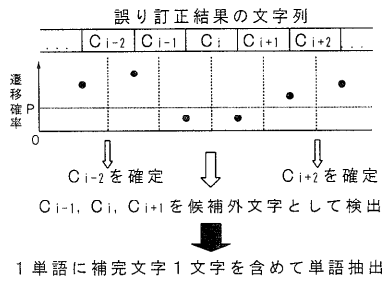


図2 Bigramによる候補外文字の検出
Fig. 2 Error detection by using bigram.

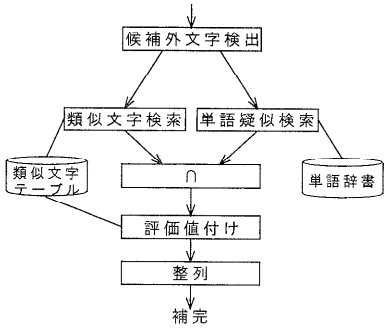


図3 補完文字生成の流れ
Fig. 3 Schematic diagram of the candidate generation process.

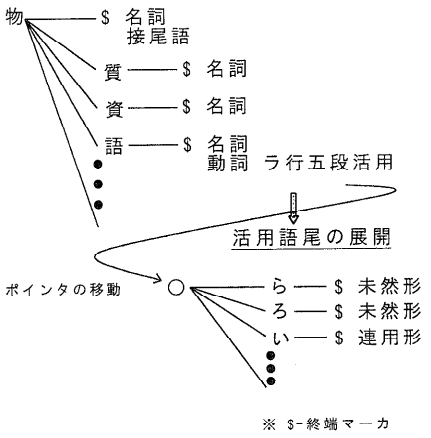


図4 単語辞書
Fig. 4 A word dictionary.

この時，下位の候補文字については考慮しない．そして，新たな候補文字を生成し，1単語に2文字以上新たに生成された候補文字が含まれないように候補文字補完，単語抽出を行う(図2)．

4. 補完文字の生成

検出された候補外文字に対して生成する新たな候補文字(補完文字)が適切であるためには，

1) 文脈に適した文字であること

2) 入力された字形パターンと無関係でないこと

という2つの条件を満たさなければならない．また，大分類をやり直した結果として誤り訂正処理，あるいは，文字認識における細分類をやり直すためには，補完文字の信頼性を示す評価値を付与する必要がある．

本誤り訂正処理手法における補完文字生成処理では，日本語に関する知識の利用として，候補外文字の前後の文字を用いて単語辞書の疑似的な検索を行い，第1の条件を満たす文脈に適した補完文字を生成する．さらに，生成された補完文字の中から，字形パターンに関する知識として文字認識部が誤って認識する可能性のある文字と，そのような誤りを犯す確率を類似文字テーブルに保持し，この情報をもとに入力された字形パターンと無関係でない文字だけを抽出し，テーブルに保持された確率をもとに評価値付けを行う(図3)．このようにして生成された言語情報および字形パターン情報の両者に支持される適切な文字により候補文字補完を行う．単語辞書の疑似検索による補完文字の生成については4.1節で，類似文字テーブルについては4.2節で述べる．

4.1 単語辞書の疑似検索

本誤り訂正処理手法の実現には，文字をノードとした木により文字列を表現した単語辞書を用いる(図4)．このような単語辞書のデータ構造はトライ(TRIE)と呼ばれる．この単語辞書は，ある単語について語頭から1文字ずつ照合を進めるごとに，次に照合可能な文字の範囲が限定されていくという特徴がある．ある文字まで検索ポインタが到達すると，次にポインタが遷移できる文字へのリンクが張られている．このリンクによりポインタの遷移先が限定される．ポインタの遷移先である文字の下は，同じくトライ構造化された部分辞書である．この再帰的な構造を利用して1単語ができる限り長い文字列で検索されるように構築されている．

単語辞書の疑似検索は候補外文字が検出された箇所にのみ適用する．検出された候補外文字の数文字前から通常の単語照合を行い，候補外文字の箇所について疑似的な照合を行う．疑似検索のアルゴリズムを1文字の事前照合を行う場合(図5の事前照合文字数＝1)の例を用いながら説明する．

現在，0個以上(例では1個)の文字について照合が行われ，次に検出された候補外文字C0，続いてC1，C2，…という順で照合を進めていくものとする．この時，以下のようなステップにより疑似的に辞書照合を

進める．ここで，ある文字Cに辞書検索ポインタが遷移し，新たな部分辞書を得ることを文字Cによる辞書の更新と呼ぶ．

step 0　現在の辞書をDICとし，S1=$\{\phi\}$, S2=$\{\phi\}$, S3=$\{\phi\}$, S4=$\{\phi\}$とする．〔例えば，「物語る」という入力に対し「物」まで照合が終わり，C0を「語」の箇所とする．ただし「語」の箇所が候補外文字であるとする．〕

step 1　DICの終端マーカを除くすべての遷移可能な文字をS1に加える．〔この時，S1={質資語…}〕

step 2　全てのkey(i)∈S1について

step 2.1　key(i)でDICを更新し，それをDIC(i)とする．〔key(i)=「語」の場合DIC(i)は図4の「語」からの部分辞書〕

step 2.2　S2(i)=$\{\phi\}$とする．

step 2.3　DIC(i)のすべての遷移可能な文字について，終端マーカであればkey(i)をS4に加え，そうでなければ文字をS2(i)に加える．〔この時，S4に「語」が加えられる．また，S2(i)={ら…る…}〕

step 2.4　key(i)とS2(i)のペアpair(key(i), S2(i))を作成しS2に加える．〔この時，S2にpair(語, {ら…る…})が加えられる．〕

step 3　すべてのpair(key(i), S2(i))∈S2について

step 3.1　C1∈S2(i)であればkey(i)をS3に加える．〔この時，S3に「語」が加えられる．〕

こうして得られたS1～S4をそれぞれHINT 1～HINT 4と呼ぶことにする．上記手続きは，C0による辞書照合を行う代わりにHINT 1およびHINT 2を用いて辞書検索ポインタを遷移させることで疑似的に辞書照合を進める．これにより，C0がどのような文字であればそこまでの照合で単語が検索されるか(HINT 4)，あるいはC1以降へ辞書照合を進められるか(HINT 3)を検査することができる．そして，得られたHINT 3, HINT 4に含まれる文字が，C0に代わる新たな候補文字となる(C0の位置への補完文字)．

図5はこのような疑似的な辞書検索により得られる単語のパターンを示している．○は実際に照合が行われた文字，●は照合が省略された文字であり，この●を埋めることができる文字の集合がHINT 3およびHINT 4として与えられるC0の位置への補完文字である．ただし，辞書照合は候補外文字が検出された箇所から事前照合文字数分だけ前の文字から始められる．また，事前照合文字数0の場合，つまり，語頭が候補外文字の場合についてのHINT 1, HINT 2は辞書の内容が変更されたときにただ1回だけ計算する．

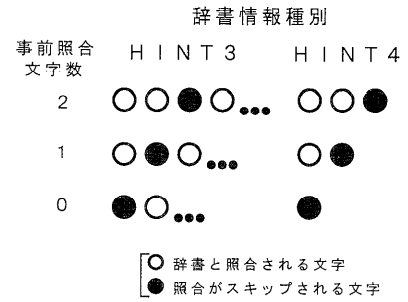


図5　疑似検索される単語のパターン
Fig. 5　Character sequence patterns obtained by pseudo retrievals of the word dictionary.

このようにして得られるC0の位置への補完文字は，文字認識部で生成された候補文字集合に正解文字が含まれない場合それを補うための有効な情報となり得る．

4.2　類似文字テーブル

上記の疑似検索は，入力された字形パターンとは無関係に新たな候補文字を生成してしまう．そこで，入力字形パターンと無関係でない適切な文字を選び出す必要がある．また，細分類のやり直しを行うためには補完文字の信頼性を示す評価値を付与する必要がある．そこで，辞書の疑似検索により生成された補完文字の中から，類似文字テーブルを用いて文字認識部から与えられた候補文字と類似した文字を選び出し，評価値付けを行う．

ここで，文字認識部へある手書きの文字C_iが入力され，認識結果としてC_jが返されたとき，C_jはC_iに類似していると呼ぶ．このような文字の類似関係について，実際の文字認識結果をもとに類似文字テーブルを作成する．類似文字テーブルには文字認識結果として与えられた各第1位の候補文字について，実際に入力された文字とそのような事象が起きる確率を記録する．これにより，ある入力に対する文字認識結果の第1位候補がC_kであるとき，正解文字として可能性のある文字はどれか，また，その文字が正解文字である確率はどれくらいであるかを知ることができる．

候補外文字として検出された文字の第1位候補文字を類似文字テーブルと照合し，類似する文字を記録された確率に従って列挙する．これを類似文字テーブルから生成された候補外文字に対する補完文字とする．この補完文字は入力された字形パターンに関する知識をもとに生成されるため，疑似的な辞書検索により生成される補完文字との間で共通する文字は，候補外文字に対する補完文字として言語情報および字形パターン情報の両者から支持される適切な補完文字であると

いえる．

5. 処理実験

本章では，言語情報にもとづく候補文字補完を用いた誤り訂正処理を実現したシステムにより，実際の文字認識結果に誤り訂正処理を適用する実験を行った．その結果をもとに，本誤り訂正処理手法の有効性について検証する．

5.1 実験方法

実験には，実際の手書き文字認識システムとしてストローク構造解析法を用いたオンライン手書き文字認識システム[8]を用いた．認識のための参照パターンには筆記者とは別の第3者のものを使用し，文字認識の結果として最大5個の候補文字を提示させた．最大5個とは，候補文字集合の大きさの最大値を5としても，文字認識システムが5個未満の候補文字しか提示しない場合があるためである．筆記者に対しては筆順や画数，書体に関する指定は行わずに自由とした．唯一筆記者に課した制約は，入力デバイスとして用いたタブレット上で，文字枠として区切られた範囲からはみ出さないように筆記することである．これにより，文字認識における文字切り出しの問題を無視する．

入力には17の入力データを用い，結果の集計のため3つにグループ化した．内訳は，文献9)から2つの節(入力文書1，2)と文献10)から7つの節(入力文書3～9)を任意に選び，入力文書1，および，入力文書2を5人の筆記者がそれぞれ筆記し，文字認識システムに認識させたデータをそれぞれ入力データ群1，入力データ群2．入力文書3～9を合計8人が筆記し，認識させたデータを入力データ群3とした．

形態素解析のための単語辞書には，実際のワードプロセッサに使用されている辞書から約3万語の自立語を流用し，見出し語として収容した．Bigramには，情報処理関係の文献から約50万文字分の文章を抜き出し，学習用テキストとして用いた．実験で用いた誤り検出のための閾値は0.01%である．類似文字テーブルの学習には，今回はハードウェア資源の制約のため，入力データのうち誤り訂正処理の対象となっているデータを除く全入力データを学習データとし，下で述べる実験1，実験3に共通に用いた．学習量は，全入力データの文字数から誤り訂正処理の対象である入力データの文字数を減算した文字数である．

本誤り訂正処理システムを適用することにより，以下に示す3種類の条件下で実験を行った．ただし，候補文字集合の大きさについて断らない場合には，最大5個の候補文字集合によるものとする．また，実験に用いた手書き文字認識システムでは「つ」と「っ」など形状がほぼ同一で大きさのみが異なる文字の識別が不可能であることがわかっているため，これらの文字については言語情報にもとづく候補文字補完とは別に，常に候補文字を補完して誤り訂正処理を行った．単語辞書疑似検索の事前照合文字数は0，1，2の3つのパターンを用いた．

実験1：言語情報にもとづく候補文字補完を用いた誤り訂正処理を適用．

実験2：言語情報にもとづく候補文字補完を用いずに誤り訂正処理を適用．

実験3：文字認識結果の第1位候補のみを用いて，言語情報にもとづく候補文字補完を用いた誤り訂正処理を適用．

実験1により，誤り訂正処理の有効性を検証し，また，入力データの違いによる誤り訂正処理の効果について検討する．さらに，実験1と実験2の結果を比較することにより，言語情報にもとづく候補文字補完の有効性について検証し，その効果について検討する．また，実験3により文字認識部が次候補を提示しない場合での，本誤り訂正処理の効果の有無について検討する．

5.2 実験結果

入力データに関する基礎的な数値を**表1**に示す．正解率は第1位の候補文字が正解であった割合，累積正解率は候補文字集合に正解が含まれた割合を示す．最大候補文字数1の場合の累積正解率は，正解率に等しい．

実験1と実験2の結果として，誤り訂正処理前の正解率と誤り訂正処理後の正解率の関係を散布図で示す(**図6**)．また，正解率の平均値を**表2**に示す．図6で×が候補文字補完を行わない従来手法による誤り訂正処理の結果であり，○が言語情報にもとづく候補文字補完を用いた誤り訂正処理の結果である．候補文字補完を行った場合の誤り訂正処理後の正解率と処理前の正解率の間には強い相関が認められ，2次曲線で近似できる(重相関係数≒1)．

表1 入力に関する基礎データ
Table 1 Properties of the input data.

	入力データ群		
	1	2	3
データ数	5	5	7
文字数	5260	6310	16363
正解率	85.6%	85.3%	81.3%
累積正解率	95.6%	94.8%	93.2%

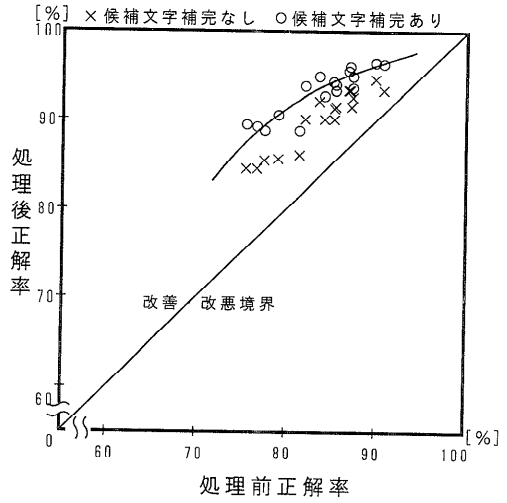


図6 誤り訂正処理の効果
Fig. 6 The effect of the error correction.

表 2 正解率の平均
Table 2 Average of the recognition success rate.

	処理前	処理後	
		(補完無し)	(補完有り)
平均正解率	83.8%	90.0%	93.1%

表 3 候補文字補完の精度
Table 3 Accuracy of the candidate supplementation.

		入力データ群		
		1	2	3
誤り検出	第1種の誤り	15.0%	13.2%	15.7%
	第2種の誤り	56.5%	47.5%	48.0%
	補完対象文字	13.6%	14.7%	17.6%
補完文字生成	補完文字正解率	90.9%	87.4%	79.5%
	類似文字正解率	93.3%	91.5%	82.7%
	疑似検索正解率	97.5%	96.0%	96.2%
	平均補完文字数	1.9	2.0	2.0
候補外文字訂正率		57.8%	54.0%	43.1%

実験1における候補文字補完の精度に関する結果を表3に示す．**表3**で，第1種の誤りは，Bigramによる誤り検出で，正しく検出されなかった誤りの，全誤りに対する割合を示す．第2種の誤りは，同じく誤り検出で，正しいにも関わらず誤りであるとして検出されてしまった文字の，誤りとして検出されたすべての文字に対する割合を示す．補完対象文字数は候補外文字として検出された文字の全文字数に対する割合を示す．補完文字正解率は，候補文字補完により正しく正解文字を補完できた割合を示す．類似文字正解率は類似文字による補完文字に正解が含まれた割合を示し，

表 4 訂正率
Table 4 Correction rates.

	入力データ群		
	1	2	3
実験1	64.0%	56.3%	54.0%
実験2	47.2%	32.7%	35.7%
実験3	48.0%	42.0%	36.8%

疑似検索正解率は辞書の疑似検索による補完文字に正解が含まれた割合を，類似文字正解率を100%としたシミュレーションにより求めたものである．ただし，これらの補完文字は文字認識システムから与えられた候補文字との重複を含む．平均補完文字数は事前照合文字数のパターンごとに，実際に補完された文字数の平均を取ったものである．候補外文字訂正率は，正しく訂正された候補外文字の割合を示している．

表4に，1から3のそれぞれの実験について，誤り訂正処理の適用により文字認識の誤りが減少した割合(訂正率)を示している．

5.3 考　　察

図6から，候補文字補完を用いるか用いないかによらず，誤り訂正処理後の正解率が処理前の正解率を上回り，言語情報を用いて文字認識結果の誤り訂正処理を行うことが有効であることがわかる．また，本誤り訂正処理手法における言語情報にもとづく候補文字補完を用いた場合，誤り訂正処理の効果がさらに増大していることがわかる．表2から，候補文字補完の導入により誤り訂正処理の適用による正解率の増分が候補文字補完を用いない場合に比べ50%増大していることがわかる．同時に，誤り訂正処理後の正解率が誤り訂正処理前の正解率に強く影響されていることがわかり，より信頼性の低い文字認識システムに対して適用する場合には，処理前の正解率の影響を小さくするための改善が必要である．

表3から，候補外文字検出のためのBigramによる誤り検出について，第1種の誤りが15%程度と大きいことがわかる．これは，本システムでは誤り訂正処理の前処理として文字の信頼性について評価を行っており，前処理の段階で誤りの可能性があるとされなかった文字については誤り訂正は行われず，また，候補外文字検出の段階においても検出されないため，誤りの総数の減少にともなって相対的に数値が膨らんでしまっているためである．また，入力データとBigramの学習データとの関係について特別の配慮を行っていないことも原因の1つである．しかし，本誤り訂正処理手法における候補外文字の検出は，特定の1文字を候

補外文字と認定することよりも，候補外文字と候補外文字の影響を受けている文字の存在範囲を限定するという意味が強く，実用上大きな問題とはならない．ただし，今後精度向上を図るという見地からは改善すべき点である．

同じく表3から，事前照合文字数のパターンごとに平均2文字により候補文字補完を行うことで80%～90%の確率で正解文字が補われていることがわかり，本誤り訂正処理手法における補完文字の生成手法が有効であったことがわかる．また，候補外文字の50%程度が正しく訂正されており，大きな成果が得られたといえる．

しかし，入力データ群3については他の入力データ群に比べ候補外文字訂正率が著しく低いことがわかる．同時に，補完文字正解率，特に類似文字正解率が他の入力データ群の場合に比べ著しく低い．この類似文字正解率の低さが候補外文字正解率に悪影響を及ぼしている．これは，類似文字テーブルの学習データの片寄りに原因があると考えられる．入力データ群1と入力データ群2は，同じ内容の文書を複数の筆記者が筆記しており，これらが類似文字テーブルの学習データに使用されているため，入力データ群1と入力データ群2については，文書中に出現するすべての文字について類似文字が学習されているが，入力データ群3については必ずしも文書中に出現するすべての文字が学習されているとは限らない．このような学習データの片寄りのために入力データ群3についての類似文字正解率が低く，候補文字補完の効果を阻害していると考えられる．

表3と表4から，誤り訂正処理により減少した文字認識の誤りの割合を，すべての誤りを対象として示した訂正率と，候補外文字だけを対象として示した候補外文字訂正率を比較すると，それほど大きな差が存在しないことがわかる．これは，本誤り訂正処理手法における候補文字補完により，誤り訂正処理における候補外文字の影響が著しく減少したことを示している．同時に，候補外文字の影響を受けない状況下で表4に示される訂正率を実現することが従来手法における誤り訂正処理，つまり，言語情報を用いた細分類の限界であることを示しており，今後，従来手法による誤り訂正処理の精度向上が必要である．

次候補を用いないで誤り訂正処理を行った実験3の結果から，本誤り訂正処理手法では，次候補を用いない場合においても誤り訂正処理により40%程度の誤りが訂正可能であることがわかり，従来手法との大きな違いが明確にされた．また，学習データの片寄りにより十分な候補文字補完の効果が上がらなかった入力データ群3を除くと，次候補を用いずに誤り訂正処理を行うことによって，実験1と実験2の場合に比べ入力データの違いによる訂正率のばらつきが減少する傾向がみらる．これは，今後システムを改善する上で注目すべき結果であった．

6. おわりに

本論文では，文字認識結果の誤り訂正処理における候補外文字の問題を解決するための手法として，言語情報にもとづく候補文字補完について述べた．本誤り訂正処理手法は，従来手法による誤り訂正結果から候補外文字を検出し，検出された候補外文字に対して日本語に関する知識と文字の字形パターンに関する知識を用いることにより，適切な候補文字を新たに生成することで候補外文字の問題に対処するものであった．

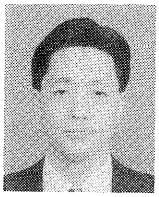
実験の結果，言語情報にもとづく候補文字補完の導入により誤り訂正処理の適用による文字認識正解率の増分が50%増大することが確認され，本誤り訂正処理手法の有効性が確認された．同時に，候補外文字の影響を受けない状況下で従来手法により候補文字集合から正解を選び出す精度の限界が明らかとなり，今後，従来手法の精度向上が課題である．また，文字認識部への情報還元についても実現したい．

参考文献

1) 西野文人：文字認識における自然言語処理，情報処理，Vol. 34, No. 10, pp. 1274-1280 (1993).
2) 池原 悟，白井 諭：単語解析プログラムによる日本文誤字の自動検出と二次マルコフモデルによる訂正候補の抽出，情報処理学会論文誌，Vol. 25, No. 2, pp. 298-305 (1984).
3) 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，情報処理学会研究報告，NL97-5, pp. 29-35 (1993).
4) 伊東伸泰：Bigramによるオンライン漢字認識の文脈後処理手法，情報処理学会研究報告，NL97-6, pp. 37-44 (1993).
5) 杉村利明：候補文字補完と言語処理による漢字認識の誤り訂正処理法，信学論(D-II)，Vol. J72-D-II, No. 7, pp. 993-1000 (1989).
6) 高尾哲康，西野文人：日本語文書リーダ後処理の実現と評価，情報処理学会論文誌，Vol. 30, No. 11, pp. 1394-1401 (1989).
7) 三部裕史，大森健児：信頼性の低い文字認識結果に対する言語情報を用いた誤認識文字の訂正，情報処理学会論文誌，Vol. 34, No. 10, pp. 2117-2124 (1993)
8) 大森健児，春木義仁：仮説設定を特徴としたオン

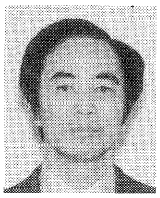
ライン手書き漢字認識，信学論(D-Ⅰ)，Vol. J74-D-Ⅰ, No. 2, pp. 130-136 (1991).
9) 小林一喜：テムズの川霧が消えた，朝日新聞社，(1991).
10) 松下幸之助：人間としての成功，PHP 研究所，(1989).

(平成 6 年 3 月 30 日受付)
(平成 7 年 1 月 12 日採録)



三部　裕史（正会員）

昭和 43 年生．平成 4 年法政大学工学部経営工学科卒業．平成 6 年法政大学大学院工学研究科システム工学専攻修士課程修了．同年富士ゼロックス株式会社入社．法政大学在学中，言語処理による文字認識後処理の研究に従事．



大森　健児（正会員）

昭和 44 年東京大学工学部計数工学科卒業．昭和 47 年カリフォルニア大学バークレイ校大学院修士課程修了．昭和 44 年日本電気入社．昭和 60 年法政大学教授．マルチプロセッサシステム，CAD 専用装置，オブジェクト指向言語，オンライン/オフライン手書き漢字認識，金型設計エキスパートシステムなどの研究に従事．工学博士．昭和 59 年情報処理学会論文賞受賞．情報処理学会 25 周年記念論文に選定．著書「システムプログラム入門」(岩波書店)，「EWS 入門」(海文堂)など．電子情報通信学会，日本ソフトウェア科学会，IEEE Computer Society 各会員．