日本語文書リーダ後処理の実現と評価†

高尾哲康†† 西野文人††

日本語文書リーダは，世の中で流通している雑誌，書籍，公文書等の印刷文書を読み取り，計算機で使用されているコード情報に変換する装置である．文書リーダ後処理は，文書リーダ装置の文字認識部で認識した結果の候補文字集合列に対して，単語照合，文法検査などの言語処理を施し，正解文字列の推定を行う．推定方式としては，(1)文字認識から得られる各候補文字に付けられた評価値，および，(2)単語照合検査，単語間接続可能性検査や文字接続確率などの言語的制約によって計算される評価値，の2つに基づく方式を考案し実現した．その結果，後処理前認識率，すなわち文字認識のみの認識率が 90% 以上の場合，後処理を施すことにより，認識率をほぼ 95% 以上に高めることが可能になった．本論文では，後処理における諸問題とその解決策，本システムの処理方式と今後の課題について述べる．

1. はじめに

文書リーダ装置は，雑誌，書籍，書類などの印刷文書を読み取り，計算機で使用されている文字コード情報に変換する装置である．オフィス等においては，文書の入力作業を自動化したいという強いニーズがあるが，従来の文書リーダは読み取り対象となる文書に制限があり，その利用は限られていた．例えば，読み取り対象文書が伝票・帳票など，1文字1枠のようにフォーマットが定まったものや住所・氏名・項目名など，記入する項目があらかじめ定まっているものに限られていた．そのため，通常オフィス等で使用されている一般文書を読み取り可能な文書リーダ装置に対する需要が高まってきている．

この装置の中心となる技術のひとつは文字認識技術である．現在の文字認識技術の状況は，認識対象の文書を英文印刷文書に限定した場合には，ほぼ実用的な95% 以上の認識率が得られている．さらにスペルチェッカを併用すれば，99% 以上の認識率を得ることもできる．しかし，日本語文書を対象とした場合には，以下の理由から実用的な認識率は得られていない．

(1) 文字数，文字種が多く，類似文字も多い．

例えば，『口』(漢字) と『ロ』(カタカナ)，『一』(長音) と『一』(漢字)，『へ』(ひらがな) と『へ』(カタカナ) など文字の形状の特徴の識別に頼る文字認識技術のみでは識別困難である．

(2) 漢字やカナには分離文字が多い．

例えば，雑誌などの文字ピッチが不定の文章を読み取る場合，横書き文章中の『読』の字は『言』と『売』の2文字と読むことができる．ワープロ文書などで使用される全角文字，または半角文字のいずれかの認識に限るならば，文字ピッチが限定されるので，文字切り出しは比較的楽であった．しかし，一般雑誌やDTP (デスクトップ・パブリッシング) により作成された文章など文字ピッチが不定の場合には，文字切り出し候補が1本にしぼれず複数になることが多い．

文書リーダ後処理はこれらの場合に対処するために，文字認識結果に対して後処理として，分離文字に対応した処理を組み込んだ単語照合，文法検査などの言語処理を施すことによって認識率を上げる技術である．この技術を導入することにより，後処理前認識率，すなわち文字認識部の1位認識率が 90% 以上 (4位以内認識率で97% 以上) あれば，後処理を施すことにより，認識率をほぼ 95% 以上に高めることが可能になった．

言語処理に基づく文字認識後処理方式は，Kawadaらによる研究[1]に始まり，いくつかの方式が提案されている．現在のところ，実行時間や技術的な面から，単語辞書との照合検査と単語間の文法的接続検査による方式[1),3)~6)]や単語辞書を使用せず，2文字連接確率を利用する方式[2)]といったレベルのものが使用されている．前者の方式の中ではさらに文字認識率の程度に応じていくつかの方式が使用されている．認識率が高くリジェクト文字 (文字認識において認識できなかった文字) が少ない場合は候補文生成方式[1),4)] (図 1) が使用されている．また，認識率が比較的低く候補文字が多い場合でも有効な方式としては探索木生成方式[2),5),6)] (図 2) が使用されている．

候補文生成方式では，リジェクト文字を含む解析範囲の決定に句読点および字種の変り目を利用し (ひら

† Implementation and Evaluation of Post-processing for Japanese Document Readers by TETSUYASU TAKAO and FUMIHITO NISHINO (Software Laboratory, Fujitsu Laboratories Ltd. Kawasaki).

†† (株)富士通研究所川崎研究所ソフトウェア研究部

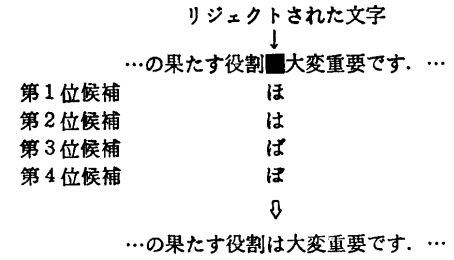

図 1 候補文生成方式
Fig. 1 Candidate sentences generation method diagram.

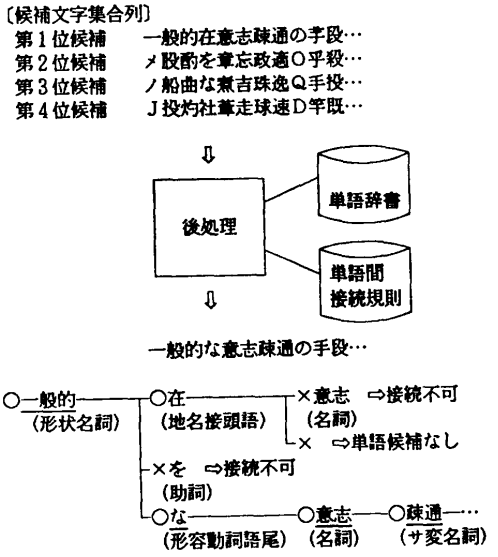

図 2 探索木生成方式
Fig. 2 Searching tree generation method diagram.

がな⇨非ひらがな，図1では『役割…です．』の間を解析範囲とする），この部分に対して形態素解析を行い，正解文字を推定する．探索木生成方式では，文頭（または文末）から各候補文字を組み合わせて構成される単語を単語辞書から検索しながら単語間の接続をチェックし，句読点まで到達した単語列を正解文字列と推定する．

いずれの方式を利用した場合でも，後処理の実用性という観点から見ると，以下の問題点を解決しなければならない．

（1） 処理速度の向上

特に，単語辞書検索の高速化が必要である．さまざまな文章を読み込んで後処理を行えるようにするためには，十分な量の語彙をもつ単語辞書が必要である．単語辞書の語彙数が増加しても単語辞書検索を十分高速に行える必要がある．

（2） 未登録語に対する対処

いくら単語辞書に含まれる単語数を増やしたとしても未登録語の出現は避けられない．未登録語が出現しても処理速度を落とさずにかつ未登録語部分およびその周囲の部分に悪影響を及ぼさないような処理が必要である．

（3） やや低い認識率でも後処理の効果があること

コピーをした書類を読み込ませた場合など，1位認識率が 90% 程度とやや低い場合でも後処理の効果が期待できる必要がある．

本システムの後処理方式は上記の探索木生成方式に基づき，上記の問題点をほぼ解決し，十分実用的な方式として実現した．

2. 後処理の処理方式

2.1 制御戦略

探索は A* アルゴリズムに基づく最良優先探索法 (best-first-search) によって行う．探索時に次に展開するノードの評価値計算には，文字認識部から得られる各候補文字に付けられる評価値と言語的制約による評価値の2つを利用する．言語的制約による評価値は，辞書に単語が存在するか否かや単語長，単語間の文法的接続可能性，単語間の接続に関する経験則などから計算される．評価値の計算は以下のようにして行う．

長さ n の入力文字列 $c_1\ c_2\ c_3 \cdots c_n$ の文字認識における観測値を，

$$\boldsymbol{X}_1\ \boldsymbol{X}_2\ \boldsymbol{X}_3 \cdots \boldsymbol{X}_n.$$

とすると，第 i 位の候補文字およびその評価値は，

$$d_{1i},\ P(\boldsymbol{X}_1|d_{1i}).$$
$$d_{2i},\ P(\boldsymbol{X}_2|d_{2i}).$$
$$d_{3i},\ P(\boldsymbol{X}_3|d_{3i}).$$
$$\vdots$$
$$d_{ni},\ P(\boldsymbol{X}_n|d_{ni}).$$

となる．ただし，$i<k$ について，

$$P(\boldsymbol{X}_j|d_{ji})>P(\boldsymbol{X}_j|d_{jk}).$$

$P(\boldsymbol{X}_j|d_{ji})$ は，文字 d_{ji} を認識したときに観測値 $\boldsymbol{X}_j$ が得られる確率である．この確率値は理想的な値なので，実験では文字認識から得られる距離値から計算により得られる擬似的な確率値を利用している．

言語的制約は以下のように計算される．

（1） 単語出現確率（頻度）

単語 W_j は各候補文字から構成される．

$$W_j=d_{11},\ d_{12},\ d_{13},\ \cdots,$$

$d_{11}d_{21}$, $d_{11}d_{22}$, …,
$d_{12}d_{21}$, $d_{12}d_{22}$, …,
….

単語出現確率は,

freq (W_j).

と表せる. freq (W_j)=0 は, 単語 W_j が単語辞書に存在しないことを意味する.

(2) 単語間接続確率

単語 W_i の次に単語 W_j が接続する確率 A_{ij} は,

$$A_{ij}=P(K(W_j)|K(W_i)).$$

ただし, $K(W_i)$ は単語 W_i の単語クラス (品詞, 活用型, 活用形, 文字種等) とする.

単語 W_i までに計算された評価値を,

$$f(W_1W_2W_3\cdots W_i).$$

とすると, 次の単語が W_j の場合の評価値は,

$$f(W_1W_2W_3\cdots W_iW_j)=f(W_1W_2W_3\cdots W_i)\times\text{freq}(W_j)\times A_{ij}\times O(W_j)+h(W_j).$$

となる. $O(W_j)$ は単語 W_j の文字認識における評価値であり, 単語 W_j を構成する各文字の文字認識における評価値 (上記の $P(X_j|d_{ji})$) から計算される. また, $h(W_j)$ は, 単語 W_j 以降, ゴール (句読点) までに計算される評価値の予想値であり, 実験では単語 W_j の単語長に基づく関数を利用している.

2.2 単語照合

単語照合は次に検索すべき単語の先頭文字をキーにして単語辞書を検索し, 検索された単語が候補文字集合列の部分集合になっているかどうかを検査することで行う. 次の2つのステップで行う.

(1) 単語検索

指定された候補文字に対して, その候補文字から始まるすべての単語を単語辞書から検索する. 図2の例では, 1文字目の1位候補文字の『一』からは『一』, 『一円』, 『一般』, 『一気』, 『一挨』, 『一般的』, …, の単語がすべて検索される.

(2) メンバ検査

(1)の単語検索によって検索された単語の表記のそれぞれの文字が, 対応する候補文字集合の要素になっているかどうかを検査する. 図2の例では, 『一』で始まる単語のリストからは『一』, 『一般』, 『一般的』が残る. また, 単語辞書内の各単語エントリは単語表記の前方一致圧縮を行っているので, 余分な検査を除くことができ, 結果として単語辞書の容量の削減と単語照合速度の向上を実現している.

このアルゴリズムでは, 単語の先頭文字が候補文字群にないときは, その単語を検索することができないが, 後処理前認識率が十分高い場合は正解文字が候補文字群に含まれない確率は十分小さいものとし, この場合を無視することにした (実験では, 単語の先頭文字が候補文字群に含まれない確率は平均で約 0.3% であった).

2.3 単語間接続検査

上記の単語照合によって得られた単語の品詞を調べ, 前の単語に接続可能かどうかを検査する. 検査には単語の品詞ごとに右接続条件, 左接続条件をもつ単語間接続規則を用いる. 図2の例では『一』, 『一般』, 『一般的』のいずれも文頭に接続可能なのでこれらの3つの単語が単語候補として得られる.

単語検索の処理速度を考えると1回の単語検索で得られる単語数が問題になる. 実験によれば, 85,000 語の単語辞書では最高で 660 個 (先頭が『大』で始まる単語), 1回の単語検索で検索される単語数の平均は約 145 個であった. 候補文字を4位までとすると, その中からメンバ検査によって残される単語数の平均は約 4.5 個, 単語間接続検査により残される単語数の平均は約3.5個であった. また, 単語辞書へのアクセスは約 1.5 文字ごとに1回あった. したがって, 1回の後処理の単位である句読点間の文字数を約 23 字[7] とすると約 2,200 単語が単語検索によって得られ, 約 70 単語がメンバ検査によって残され, さらに, 約 54 単語が単語間接続検査によって残される. 実際には1回の単語辞書アクセスごとに, 後述の未登録語テンプレートがメンバ検査によって残された単語候補ごとに加わる. 未登録語テンプレートの数を 20 個とすると接続検査によって残されるテンプレート数の平均は約 2.2 個であった. したがって, 探索木の大きさの平均は, 単語候補数にして,

$$(3.5+2.2)\times 23/1.5=87.4 \text{ 個}.$$

となる.

一方, 別の単語照合方式として, 候補文字を組み合わせてできる単語が単語辞書に存在するかどうかを単語検索によりチェックする方式の場合, チェックしなければならない単語数は, $\sum_{j=1,23}\sum_{k=1,j}n^k$ 個 (n: 候補文字数, k: 単語長) と 10^{13} のオーダとなるのでほとんど実用的でない. 前述の探索木の大きさでは, 十分実用的な処理速度 (15 MIPS の計算機で 1.0〜1.5 msec/字) および必要メモリ量 (作業用メモリ量 256

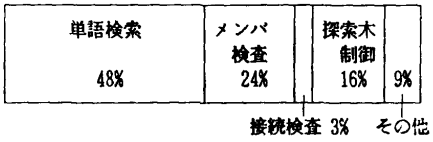

図 3 各部分の実行時間に占める割合
Fig. 3 Rate of execution time of each part of post-processing program.

KB 以下）を実現できた．図3に現在の後処理の各部分の実行時間が全実行時間に占める割合を示す．

3. 後処理における諸問題と解決策

3.1 単語表記のゆれ

機械翻訳システムでは前編集時に，自然言語インタフェースシステムでは問い合わせ文の入力時に，単語をなるべく漢字で書かせ，送り仮名表記や異表記を統一させるような制限を加えることは可能であった．しかし，文書リーダではすでに書かれている文章を読み込まなければならないため，前編集などは行えず，このような制限を読み込み対象の文書に設けることは望ましくない．単語表記のゆれには以下のようなものがある．

（1） 送りがな表記のゆれ（例.『行なう』と『行う』）

（2） ひらがな表記（例.『推敲』と『推こう』）

（3） 外来語カタカナ表記のゆれ（例.『インタフェース』と『インターフェイス』）

（4） 異表記（例.『国』と『國』）

これらに対処する方法としては，あらゆる可能性を辞書に登録する方法，プログラムで処理する方法[8]が考えられるが，前者の方法が有用であると考える．その理由は，文書リーダ後処理用の単語辞書は機械翻訳用の単語辞書と違い，ひとつの単語エントリにもたせる情報が単語表記と接続情報のみと少ないこと，後処理プログラムが複雑になることを防ぎうるからである．さらに，送りがな表記のゆれをプログラムで処理する方式を実現した中村[8]の方法では必要以上の送りがな表記のゆれを吸収してしまうことが実験によりわかっていることも要因の1つである．

我々は，(1)送りがな表記のゆれに対しては約 500 語，(2)ひらがな表記については和語形容詞，和語動詞，常用外漢字をもつ単語を中心にそのひらがな語を約 5,000 語，(3)外来語カタカナ表記のゆれと(4)異表記については約 1,000 語，の登録を行うことで解決した．

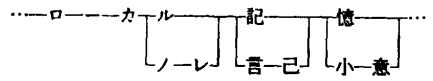

図 4 分離文字の例
Fig. 4 An example of separable character.

3.2 分離文字

日本語文字には，へん＋つくりなどの部首で構成される漢字や『は』,『ル』のように分離部分のあるカナが多く含まれている．また，多くの文書には英数字等の半角文字も含まれている．このため，文字認識では分離した文字や半角文字を認識するときには分離した場合の認識結果とともにそれらを統合した場合の認識結果を出力し，後処理において正解文字を確定することにした．図4に分離文字の例を示す．我々の実験によれば，分離文字として認識される文字数の割合は約 3% であった（図7参照）．

3.3 未登録語

単語照合検査を単語辞書を利用して行う限り，いかに単語辞書に含まれる単語数を増やしたとしても，未登録語の出現は避けがたい．また，文書リーダの場合は機械翻訳，自然言語インタフェースなどの場合と違い，どんな種類の文章でも読み込めることが要望されている．

文書リーダ後処理の未登録語処理は，(1)どの部分が未登録語であるかを認定する単語範囲認定部と(2)その単語を構成する文字を認定する文字認定部の2つのステップから成る．

3.3.1 単語範囲認定部

未登録語の範囲を認定する方式としては，形態素解析が失敗した時点で，その前後を解析し直して未登録語の範囲を決定する方法が提案されている[9]．しかし，この方法では，未登録語処理用の特殊な処理が必要になる．文書リーダ後処理では，各文字に対して候補文字が複数あるため，そのすべての組み合わせを調べて失敗したときはじめて未登録語であると認定する方法ではコストがかかる．また，低い評価値をもつ候補文字を組み合わせてできる単語を誤って確定する危険がある．

そこで，我々は未登録語テンプレートを利用した方法を提案した[6],[10]．未登録語テンプレートはパターンマッチが可能な単語であり，単語照合時に単語辞書から得られた単語候補と全く同等に扱われる．単語照合時に単語辞書から検索された単語候補に加えて候補文字にマッチする未登録語テンプレートが単語候補リ

ストにつけ加えられる．後処理における未登録語テンプレートの役割は，未登録語の抽出だけでなく，正解文字が候補文字にない場合や低い評価値の場合に後処理誤りが周囲に拡散することを抑えることである．

一般に未登録語のパターンは，『$\alpha\beta^*\gamma$』(α：語頭，β：語中，γ：語尾のパターン，* は0回以上の繰り返しを表す）の形をしているため，

	左接続条件	右接続条件
α：	文頭，名詞，…	β，γ
β：	α	β，γ
γ：	α，β	文末，格助詞，…

の形式をもつ未登録語テンプレートを用意しておけば，未登録語となる単語の文字数（2文字以上の場合．1文字の未登録語の場合には別に評価値を低くしたテンプレートを用意する）に関係なく未登録語の範囲決定が行える．

各テンプレートは以下の情報をもつ．

（1） マッチする文字のパターン

候補文字集合中のマッチする文字のパターンを示す．例えば，語頭カタカナ文字（カタカナ文字のうち，『ー』，『ェ』，『ン』などの語頭禁則文字を除いたもの）をαにもたせたり，固有名詞などに接続しやすいパターン（例えば，『〈未登録語〉市』）をもたせたりすることができる．

（2） 接続情報

テンプレートの前後の単語との接続条件を示す．

（3） 評価値

そのテンプレートが使用されたときの評価値であり，候補文字の評価値と同等に扱われる．

3.3.2 文字認定部

未登録語となった部分に対しては，各候補文字を組み合わせていかに単語らしい文字列を決定するかが重要である．そこで，未登録語となった部分について，文字認識部によって付けられた評価値および文字連接確率を利用することにより未登録語を構成する各文字を決定することにした[11]．

文字認定部では前項の未登録語テンプレートにより決定した単語範囲および文字種内で，文字連接確率により最終的な正解文字が推定される．

3.4 正解文字が候補文字に含まれていない場合

出現頻度が低い文字を読み込んだ場合，印字の品質が悪く文字につぶれやかすれがある場合などは，正解文字が候補文字に含まれない場合が生じる．このような場合に正解文字を推定する方式としては以下の2つを実現した．

（1） 単語部分照合処理

単語辞書に正解候補の単語が登録されていることを仮定すれば，単語照合処理において，その単語のある文字が候補文字集合中にない場合には単語候補からはずすのではなく，ペナルティを付けて単語評価値を下げておくにとどめる．この方法は，短い単語に対しては効果がなく，改悪の危険が高いため，専門用語のような比較的長い単語に対してのみ行うことにした（効果については後述）．この単語部分照合処理は，単語照合時の処理の増大や探索木の拡大により，処理時間は導入しない場合に比べ，約 3〜4 割増になった．

また，先頭文字が候補文字群にない場合の単語検索については，2文字目以降の文字をキーにして検索を行うためのインデックスを利用することで実現可能になるが，現在のシステムには実装していない．

（2） 未登録語テンプレートによる方法[6]

特に活用語尾に相当する文字が候補文字集合にない場合は，それらの品詞に対応する接続規則をもつ未登録語テンプレートを用意しておくことで救済可能なことがある．現在のシステムでは，あらゆる場合に対処するために多くの未登録語テンプレートを用意することは処理コストの増加をまねく．これを避けるため，より効率のよいアルゴリズムが必要になる．

（1），（2）のいずれの場合も，精度を高めるためには後処理によって推定した候補文字（複数）を文字認識部にフィードバックして，どの候補文字がより確からしいかを決定するためのインタフェースが必要となり，今後の課題である．

4. 実験結果

科学技術分野の，雑誌および入門用書籍を対象として認識および後処理の実験を行った．約 50 件のデータ（累計約5万字）の後処理結果を**図5**に示す．さらに後処理前認識率の 1% 刻み単位に平均をとったものを**図6**に示す．これにより，ほとんどのデータについて，後処理前1位認識率を，後処理を施すことにより，認識率をほぼ 95% 以上に向上させることができた．なお，認識率の計算は以下のようにして求めた．

正解文字数／入力文書中に出現する文字数×100.

（後処理前認識率において，分離文字の場合は統合された文字の第1位候補文字が正解のときに限りその文字を正解文字数に含めた．）

また，**図7**に未登録語部分，正解候補がない部分，

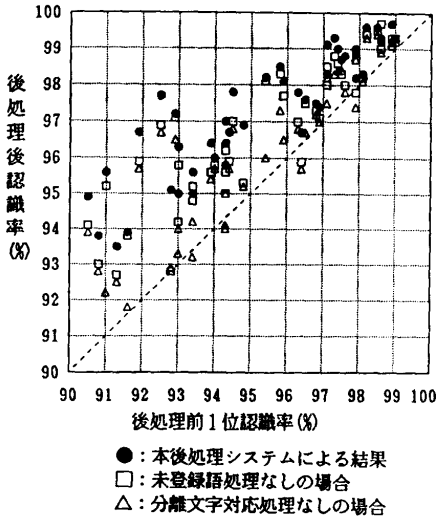

図 5 実験結果（1）
Fig. 5 Result of experiment (1) (Correction rate against recognition rate).

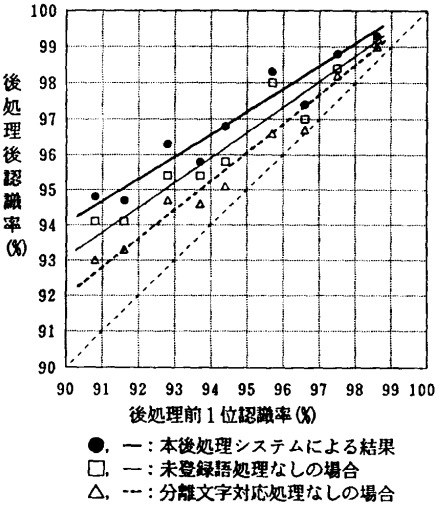

図 6 実験結果（2）
Fig. 6 Result of experiment (2) (Correction rate against recognition rate).

分離文字部分の全文書中に占める割合を示す．図6と図7から以下のことがわかる．

（1） 分離文字対応処理については，約 1% の認識率向上に寄与することが確認された．文書中の分離文字部分の割合は 2.9% なので，分離文字部分の約 1/3 が救済できた．

（2） 未登録語処理においては，約 0.5% の認識率向上に寄与した．未登録語部分もそうでない部分も後処理前認識率は違わないことを考えると，未登録語処理なしの場合，0.9×(1−後処理前認識率) だけ認識率

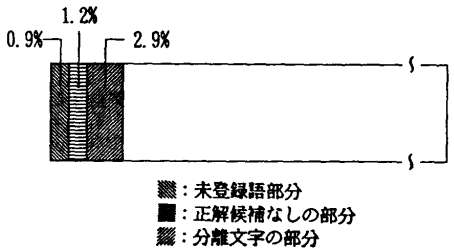

図 7 未登録語部分，正解候補なしの部分，分離文字部分の全文書中に占める割合
Fig. 7 Proportion of the parts of unregistered words, the parts in which exact character is not the member of candidates and the parts of separable characters to the whole documents.

が低下すると考えられるのに，それ以上に低下したことから，未登録語処理なしのときはその部分およびその周辺部分を改悪することが多いことがわかる．

（3） 単語部分照合処理については長い単語についてのみ行ったため，正解候補なしの部分 1.2% のうち約 4% を救済できたにとどまった．

改悪を含む救済失敗の例を**図8**に示す．この結果から以下のことがわかる．

（1） 正解文字が候補文字集合中にあるのに誤った単語を確定する場合は，ひらがな単語の語頭または語尾に付属語相当の単語が現れる場合や表記，品詞いずれも似ている単語の場合に多かった．また，後処理以外の原因による救済失敗には，文字認識における文字切り出し失敗により，単語長がはっきりしない場合によるものがあった．

（2） 分離文字の場合は長さ2以上の単語の一部の文字が分離文字の場合は容易に救済できたが，1字単語の場合は救済失敗する割合がかなり高くなった．

（3） 未登録語処理による救済失敗には，単語範囲認定部における失敗と文字認定部における失敗がある．前者は，1字単語が現れる場合が多かった．後者は，『エ』と『ェ』など同じ文字種間で形状が似ているものが多かった．

（4） 正解文字が候補文字集合中にない場合は，正解文字を含む単語の品詞により結果が分かれた．名詞，固有名詞などの場合は1字単語以外は未登録語テンプレートにより，改悪を最小限に抑えることができたが，動詞の語幹や活用語尾の場合は誤りを周囲に拡散することが多かった．この場合は用言の語幹や活用語尾に相当する未登録語テンプレートを用意することで改悪を抑えることができる[6]．

(1)単語照合　この一こと一は一,　(正解)
　　この一ことば一,　(失敗)

他に,「リンク」と「リング」,「大学」と「大字」,「間」と「問」など

(2)分離文字　起動一用一に┌ω┐一チャネルーを
　　└6－4┘

(3)未登録語抽出　ジッターとは一,　(正解)
　　ジッタヒーは一,　(失敗)

　　クリエイター一たち　(正解)
　　クリエイター一たち　(失敗)

(4)正解候補なし　会社一を一採一ろ一こと一…
　　⇧
　　正解文字「作」(動詞語幹)が候補にないので,「る」(語尾)を「ろ」(名詞)と改悪した.

図 8 救済失敗の例
Fig. 8 Some examples of failures in post-processing.

5. 考　察

前章の実験結果により，単語表記と接続規則をもとにした後処理はこれが限界と考える．今後，言語的制約として単語出現頻度情報，単語品詞間接続可能性の情報や分野別の出現情報を導入することにより，救済失敗のいくらかは救済できる．特に救済失敗が目立った1字単語に関して有効であると考える．現在の後処理用単語辞書中には，1字単語は記号，欧字，活用語尾等を除くと約2,700語ある．1字単語は単語照合と接続検査のみでは十分な言語的制約とならず，また，未登録語の範囲決定に悪影響を与えることが多い．そこで，1字漢字の接辞語（接頭語，接尾語）をとる単語はなるべく接辞語を含めた複合語として単語辞書に登録することにより，改悪をいくらか抑えることができた．1字単語の名詞（約 1,400 語）については今後，前述の制約の導入とともに類似文字の調査・分類[12]を行うことにより対処することを考えている．

さらに，人間による後処理の実験[13]によると，関連語情報，意味情報，文脈情報などを利用することにより，後処理前認識率が悪くても十分な救済能力をもつことが可能である．しかし，現時点では認識率 100% めざして大変な努力をすることは，あまり得策とは言えないので，今後の方向としては，マンマシン・インタフェースの向上など，視点を人間の側に移すべきであろう．

6. 今後の課題

6.1 誤りらしい箇所の指摘機能

認識率の向上は使いやすさの重要な要因である．しかし，完全に 100% の認識率は現状では望めないので，文書の読み込み後の人間によるチェックは欠かせない．たとえ 99.9% の認識率であったとしても全文書中 0.1% の認識誤り箇所を修正するのに全文書を見直すとなると大変である．後処理においては，候補文字評価値が低かったところや未登録語と判定した部分などを誤りらしい箇所としてユーザに指摘する機能が望まれる．

誤り指摘機能を精度よくするためには，未登録語抽出精度の向上，評価値計算の精度向上が必要である．未登録語抽出精度の向上には，未登録語の語構成パターン[14]を調査し，その結果を未登録語テンプレートに反映させていくことが必要となる．評価値計算の精度向上には，確率的手法[6),15)]が有効である．

また，誤り指摘機能はユーザによる許容認識率のレベルによって有効かどうかが決まると考える．例えば，外部へ公開する場合などは全文をチェックしてでも見直す必要がある．一方，グループ内の資料や読み込んだ文書を推敲するなど手を加えることが前提の場合には誤り指摘のレベルが緩くても実用上は問題ないと考える．

6.2 単語登録

未登録語が効率よく単語辞書に登録できる機能が望まれる．未登録語は特に専門用語の場合は同一文書中に通常繰り返し現れる．したがって，後処理が未登録語と判定した部分の単語と品詞を表示し，ユーザに確認して登録できるようにすることが望ましい．さらにその未登録語が出現した部分以降の後処理を再度試みることが必要になる．

6.3 様々な種類の文書の読み取り

本後処理は読み取り対象文書が日本語標準語の文法に従った文書でないと効果がでない．ユーザには，和英混在の文書，名簿，登記簿，裁判の判例（漢字カタカナ文）や古文などを入力したいというニーズがある．これらに対応するためには，住所・氏名辞書，専門用語辞書などの分野別辞書，文体に応じた単語間接続規則の整備などが必要となる．

謝辞 日頃，御指導いただいている杉山主任研究員，およびソフトウェア研究部第1研究室の方々に深謝致します．

参考文献

1) Kawada, T., Amano, S. and Sakaki, K.: Linguistic Error Correction of Japanese Sentences, *COLING 80*, pp. 257-261 (1980).
2) 杉村，斉藤：文字連接情報を用いた読み取り不能文字の判定処理―文字認識への応用―，電子通信学会論文誌，Vol. J68-D, No. 1, pp. 64-71 (1985).
3) 池田，大田，上野：手書き原稿認識における語彙および構文の検定，情報処理学会論文誌，Vol. 26, No. 5, pp. 862-869 (1985).
4) 新谷，目黒，梅田：言語情報と認識情報の複合利用による文字認識後処理，研究実用化報告，Vol. 34, No. 1, pp. 73-83 (1985).
5) 黒沢：日本語文書を対象とする文字認識後処理方式，第36回情報処理学会全国大会論文集，pp. 1801-1802 (1988).
6) 西野，高尾：日本語文書リーダ後処理の実現，情報処理学会自然言語処理研究会資料，64-6 (1987).
7) 林ほか：図説日本語，角川書店 (1982).
8) 中村：和文形態素解析における異なった送り仮名の自動検出と自動修正，第34回情報処理学会全国大会論文集，pp. 1297-1298 (1987).
9) 長瀬：ATLAS II における未登録語の抽出とその扱い，第36回情報処理学会全国大会論文集，pp. 1271-1272 (1988).
10) 西野，高尾：日本語文書リーダ後処理における未登録語処理，第37回情報処理学会全国大会論文集，pp. 1024-1025 (1988).
11) 高尾，西野：日本語文書リーダ後処理におけるヒューリスティック規則について，第36回情報処理学会全国大会論文集，pp. 1313-1314 (1988).
12) 梅田：単語辞書を用いた文字認識における文字の確定能力，電子情報通信学会論文誌，Vol. J72-D-II, No. 1, pp. 22-31 (1989).
13) 西野：文字認識後処理の可能性，情報処理学会自然言語処理研究会資料，62-10 (1987).
14) 亀田，森田，倉島，藤崎：未知語の分類とその処理規則，第36回情報処理学会全国大会論文集，pp. 1195-1196 (1988).
15) 長田，牧野，日高：日本語の文脈情報を用いた文字認識，電子通信学会論文誌，Vol. J67-D, No. 4, pp. 520-527 (1984).

(平成元年3月27日受付)
(平成元年7月18日採録)

高尾 哲康（正会員）

1957年生．1979年東京工業大学理学部応用物理学科卒業．1981年同大学院修士課程修了．同年(株)富士通研究所入社．1989年(株)日本電子化辞書研究所出向．自然言語処理の研究・開発に従事．

西野 文人（正会員）

1956年生．1979年東京工業大学理学部情報科学科卒業．1981年同大学院修士課程修了．同年(株)富士通研究所入社．機械翻訳をはじめとする自然言語処理の研究・開発に従事．ACM 会員．