卒業論文

単語情報を利用する
文字認識手法に関する研究

東北大学工学部　情報工学科
阿曽研究室　Ｂ４

秋山　秀三

目 次

1　序論

1.1　研究の背景及び目的

近年、情報処理技術として、計算機による知識情報処理が広まってきている。

その中で、文字認識システムは、日本語文書データの入力自動化を目指すものとして注目されており、その実現のため、認識の高精度化に関する研究が行われている。

しかし、日本語文書における類似文字、あるいは印刷文字のかすれやつぶれ、にじみといった、低品質文字については、１文字毎の文字パターンからだけでは正確な認識にも限界があると考えられる。このような文字については、入力文字列に関する知識を利用した後処理を施す事が認識精度を上げるうえで有効である。

そこで本研究では、文字認識システムにより得られた認識結果に対し、単語、文節、文法といった日本語の言語情報を活用する事により、最も確からしい文字列を決定するための手法について検討する。本研究の大まかな流れを、図1に示す。

1.2　本論文の構成

第1章は序論であり、研究の背景及び目的を述べる
第2章は本実験で用いる言語情報の形式について述べる。
第3章では本研究における単語照合アルゴリズムについて説明する。
第4章では実際に認識の後処理実験を行なう。
第5章では、実験結果の具体例をみながら考察を行なう。
第6章で、本論文の結論を述べる。

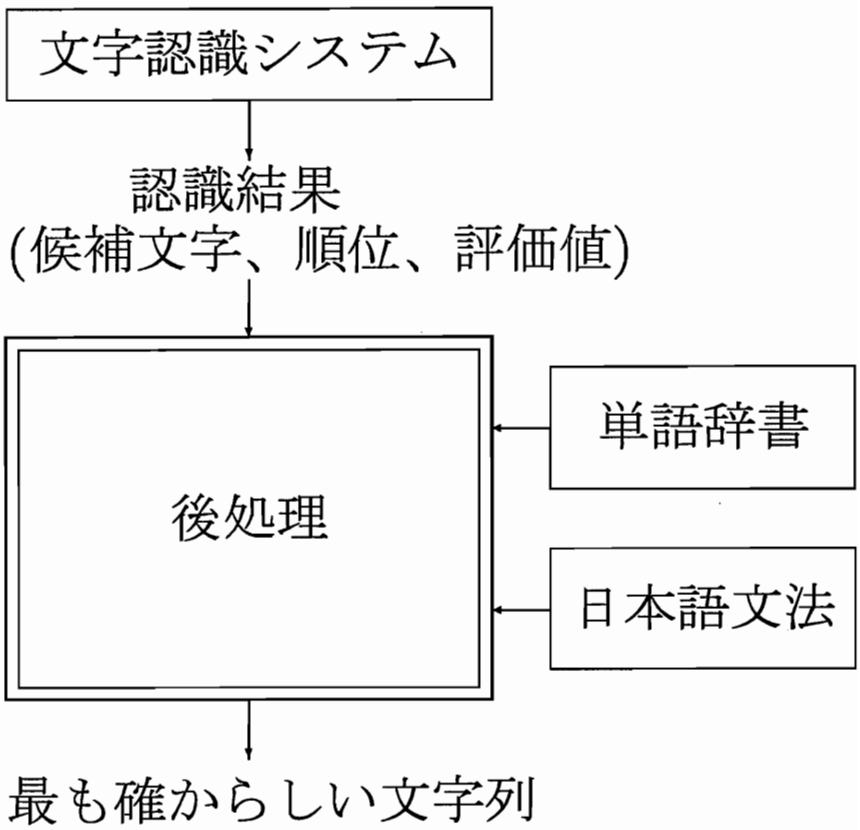

図 1: 後処理過程

2 使用する日本語知識の形式

2.1 使用する日本語知識について

本研究で使用する日本語知識として、自立語辞書、付属語辞書、そして、日本語文法データを用意した。

このうち、自立語辞書は、九州芸工大自立語辞書ＫＩＤ－Ｊ８２を基とした。また、付属語辞書は、その付録として掲載されていたものを使用した。

```
A0162 ア              1 3967 * *D1J4T S K1A2Y O H & !  % R @ N ## 合
A0163 ア              1 3271 * *E1J4T S K1A2Y O H & !  % R @ N ## 会
A0164 ア              1 3029 * *F J4T S K1A2Y O H & !  % R @ N ## 逢
A0002 アークトウ      132 00004574 * * 1J T S K Y O H & !  % R @ NC1# □灯
A0003 アーケート゛    16 * * J T S K Y O H & !  % R @ NB-#
08250 アーテイスト    1312 00002523 *1* J T S K Y O H & !  % R @ N4-O □ィ
A0225 アートシ        131 00003B66 * * J T S K Y O H & !  % R @ NC1# □紙
08155 アアマデ        23 00004B78 * * J T S K Y O H1 & !  % R1@ N ## □迄
A0011 アイ            2 3026 * *C1J T1S1K13 Y O H & !  % R @ N11# 愛
A0014 アイアール      123 23492352 * * J T S K Y O H & !  % R @ N1-# ＩＲ
Z7613 アイアラソ      23 416A4168 * * J4T S K1A Y O H & !  % R @ N ## 相争
08169 アイインシヤ    222 3026307B3C54 *1* J T S K Y O H & !  % R @ N41# 愛飲者
Z7611 アイウ          21 416A4247 * * J4T S K14 Y O H & !  % R @ N ## 相打
A0023 アイエンカ      221 3026316C3248 * * J T S K Y O H & !  % R @ N41# 1 愛煙家
A0027 アイカキ゛      23 39673830 * * 1J T S K Y O H & !  % R @ NC1# 合鍵
08237 アイキユウ      23 23492351 *1* J T S K Y O H & !  % R @ N1-# ＩＱ
Z7605 アイシア        211 302600003967 * * J4T S K1A Y O H & !  % R @ N ## 愛□合
```

※ 空白が削除されている部分有り

図 2: 九州芸工大自立語辞書ＫＩＤ－Ｊ８２(一部抜粋)

2.2 自立語辞書

「九州芸工大自立語辞書ＫＩＤ－Ｊ８２」より、必要な情報を取り出し、以下のように再構成し、漢字コード順にソートしたものを使用した。約９００００語の自立語が登録されている。

1	2	3	4	5	6	
サ変動詞	活用型	活用行	形容詞	形容動詞	その他の品詞	単語の表記

図 3: 自立語辞書の format

１．サ変動詞表示部

サ変動詞を、口語と文語*、また、清音と濁音で区別して表わした。

[1]　清音の口語・文語サ変動詞　　[愛(する)]
[2]　濁音の口語・文語サ変動詞　　[信(ずる)]
[3]　清音の文語サ変動詞　　[解(す)]

２．活用の型表示部

サ変以外の動詞を次のように分類して表した。

[1]　五段(四段)活用
[2]　上一段活用
[3]　下一段活用
[9]　カ行変格活用

３．活用行表示部

活用の型に対して、その活用行を次のように表示した。

[2]　カ行　　[9]　ラ行
[3]　サ行　　[A]　ワ(ア)行
[4]　タ行　　[B]　ガ行
[5]　ナ行　　[C]　ザ行
[6]　ハ行　　[D]　ダ行
[7]　マ行　　[E]　バ行
[8]　ヤ行

４．形容詞の型表示部

形容詞を次のように活用の型で分類して表した。

[1]　口語形容詞
[2]　文語形容詞ク活用
[3]　文語形容詞シク活用
[4]　口語形容詞と文語形容詞ク活用を兼ねたもの
[B]　口語形容詞と文語形容詞シク活用を兼ねたもの

５．形容動詞の型表示部

形容動詞を次のように活用の型で分類して表した。

[1]　口語形容動詞
[2]　口語形容動詞のうち、「～の」に形のあるもの
[3]　文語形容動詞タリ活用
[4]　文語形容動詞ナリ活用
[B]　口語形容動詞と文語形容動詞ナリ活用を兼ねたもの

*本研究では、文語文法については考慮していない。

６．その他の品詞の表示部

上記以外の品詞 (活用のないもの) を次のように分類した。

[1] 名詞
[2] 連体詞
[3] 副詞
[4] 接続詞
[5] 感動詞

2.3 付属語辞書

形式名詞、助動詞、助詞、補助用言、活用語尾の順に、５４０語の付属語が、それぞれの活用型、品詞、活用形がコード化され、リストされている。

こと	00 001 001 0	まで	00 243 243 0
ところ	00 001 001 0	やら	00 244 244 0
はず	00 001 001 0	こそ	00 245 245 0
まま	00 001 001 0	さえ	00 246 246 0
もの	00 001 001 0	しか	00 247 247 0
よう	00 001 001 0	でも	00 248 248 0
ごとく	00 054 053 2	なんて	00 249 249 0
ごとき	00 054 054 4	な	05 017 010 1
させ	00 057 055 1	の	05 017 011 1
させ	00 057 056 2	に	05 017 014 2
させる	00 057 057 3	ん	05 017 015 2
させる	00 057 058 4	ぬ	05 017 017 3
させれ	00 057 059 5	ぬ	05 017 018 4
させろ	00 057 060 6	ね	05 017 019 5
させよ	00 057 061 6	ね	05 017 020 6

図 4: 付属語辞書 (一部抜粋)

2.4 日本語文法データ

ここで取り上げた文法は、文節を構成する際の語と語の接続規則であり、自立語、付属語において、いかなる品詞にいかなる付属語が接続可能であるかを前述の付属語辞書のアドレスにより指定し、示した。また、その語で終了した場合に、文節として成立するかどうかについても示してある。

3 単語照合アルゴリズム

3.1 リジェクト文字の決定

入力された評価値より、認識部で誤認識した可能性があると思われる、つまり後処理の対象とする文字を決定する。

利用する評価値は、各候補文字に対する距離値であり、ここで用いる距離 d は、以下の計算により求めたものである。特徴ベクトルとしては、本研究室の文字認識に使用している方向線素特徴ベクトルを使用した。本研究室で行なう認識では、ここで重み付き距離を用いるが、今回は後処理の効果をみるため、重みははずした。

$$d = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{u}_i)^2$$

$\boldsymbol{x}$: 未知入力文字特徴ベクトル
$\boldsymbol{u}$: 標準パタンベクトル

リジェクト判定基準としては、1位候補と2位候補の距離比を用い、その距離比が基準値より小さい、すなわち

$$\frac{d_2}{d_1} > T \ (T = const.)$$

を満たす文字をリジェクト文字として判定する。

3.2 照合範囲の決定 (文節の切り出し)

入力された文字の中から、照合を開始する位置及び終了する位置を決定する。この位置は、文節の切れ目である事が望ましいため、句読点及び平仮名から平仮名でない文字へ文字種が変化する点で範囲をとる。その範囲にリジェクト文字を含む場合、その複数候補文字により組み合わせをつくり、次の単語照合の過程への入力とする。

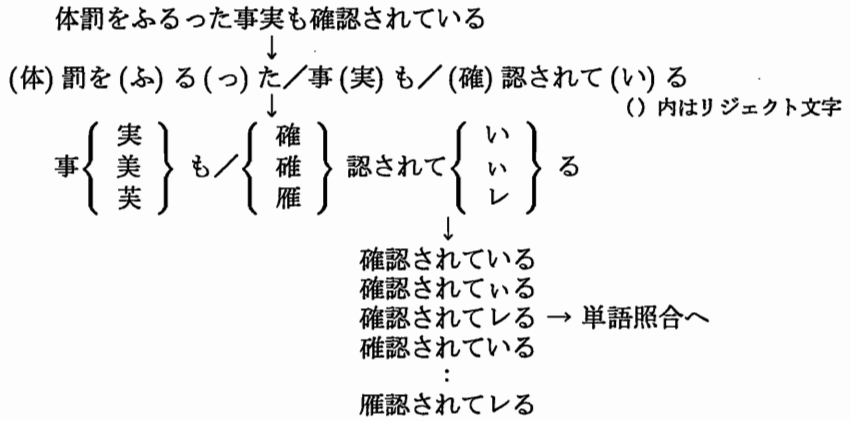

図 5: 文節の切り出し

3.3 単語照合

3.3.1 自立語照合

入力された文節と自立語辞書とを照合し、照合に成功したものについて単語候補として文節処理を行なう。自立語辞書には、活用語はその語幹しか登録されてないため、活用語である場合には、辞書内のデータよりその活用語尾をチェックする。

3.3.2 文節処理 (単語の拡張)

自立語照合により選出された単語の品詞・活用型・活用形から、その後にいかなる付属語が接続可能であるかを日本語文法データにより調べ、その付属語について辞書と切り出された文節とを照合する。照合に成功した場合、接続した付属語の状態 (品詞その他) が新たにその単語列の状態となる。接続した付属語の後に、更に付属語が接続する場合があるので、この処理を付属語が接続しなくなるまで続ける。この処理で単語がどのように拡張されるかを、図 7に示す。

この過程でつくられる単語列に関して、その時点で文節として成立するかどうかを日本語文法データにより調べ、成立する場合にはその単語列を文節候補として保持する。すべての組み合わせの入力文節についてここまでの処理が終了したら、文節候補のうちで最長のものを構成する候補文字を出力する。これが入力文節のすべてでない場合は、出力されなかった部分を新たな入力文節として再度単語照合過程へ入力する。

この処理において文節候補がない場合、また、自立語照合において単語候補がない場合には、照合範囲を 1 文字移動、つまり最初の 1 文字をのぞいた文字列を新たに入力文節とし、単語照合を行う。ここで、最初の 1 文字がリジェクト文字の場合、後処理における検証は出来なかったとして、出力としては認識部における 1 位候補を用いた。この単語照合アルゴリズムを、図 8に示す。また、「体罰をふるった事実も確認されている」という文章を入力した時の出力は次のようになる。

体罰をふるった事実も確認されている

↓

(体) 罰を/(?ふ)/る/(?っ)/た/事 (実) も/(確) 認されて (い) る

[( ) 内はリジェクト文字
// は認識された文節の範囲
? は文節として認識できず 1 位候補を出力したもの]

※『ふるう』という動詞が自立語辞書に登録されてないため、
「ふるった」が文節として認識できなかった。

図 6: 出力例

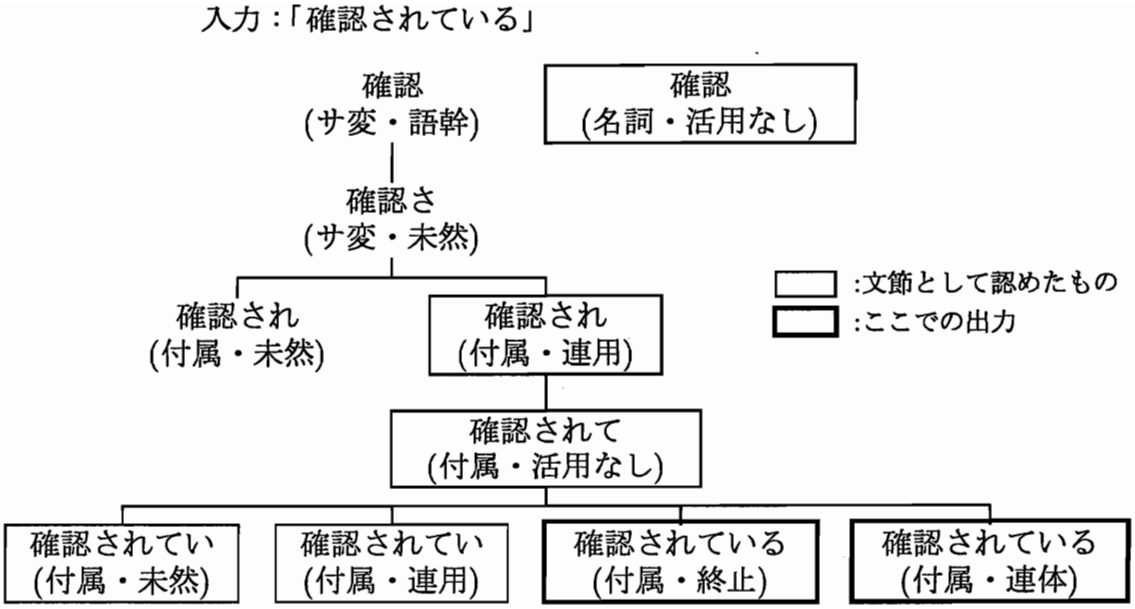

図 7: 単語の拡張例

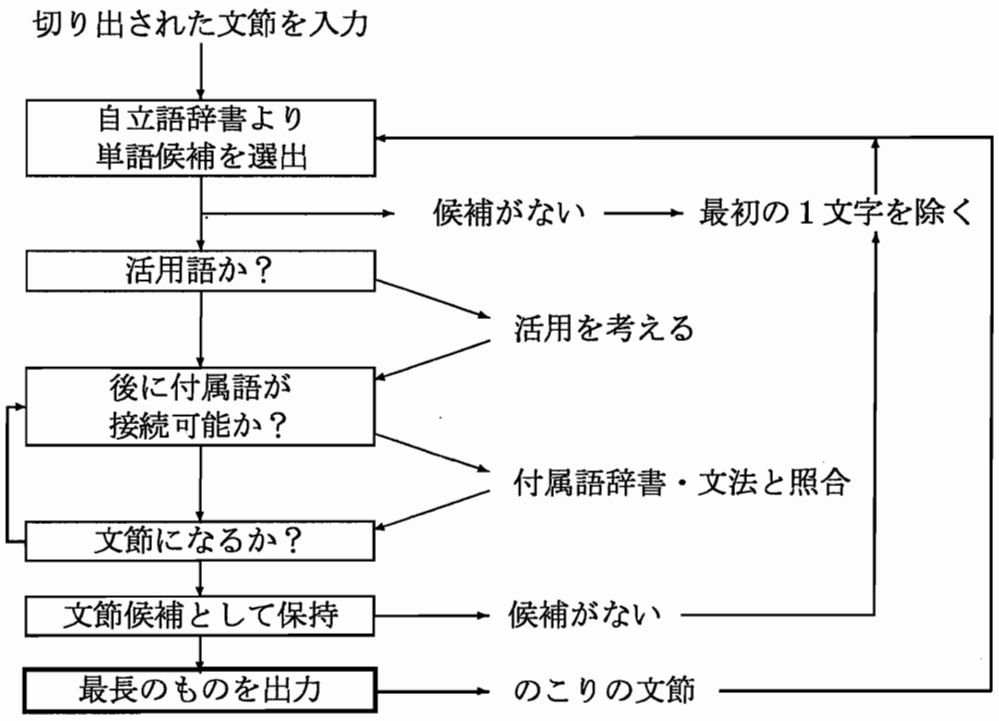

図 8: 単語照合アルゴリズム

4 実験

4.1 予備実験

後処理実験を行うにあたり、種々の値を定めるため、次の実験を行った。

4.1.1 リジェクト判定基準

リジェクト文字を決める判定基準として、1位候補と2位候補の距離比を用いるが、その閾値をどこに定めればよいかを調べるため、後処理実験で使用する文書について、距離比による後処理前の1位認識率及びリジェクトされる割合を調べた。(図 9) 閾値を大きくとれば、認識精度は高くなるが、当然、リジェクトされる文字数も多くなり、処理の効率は悪くなる。

今回の実験では、1位認識率が一定になり始める 1.7 を、閾値として用いる。

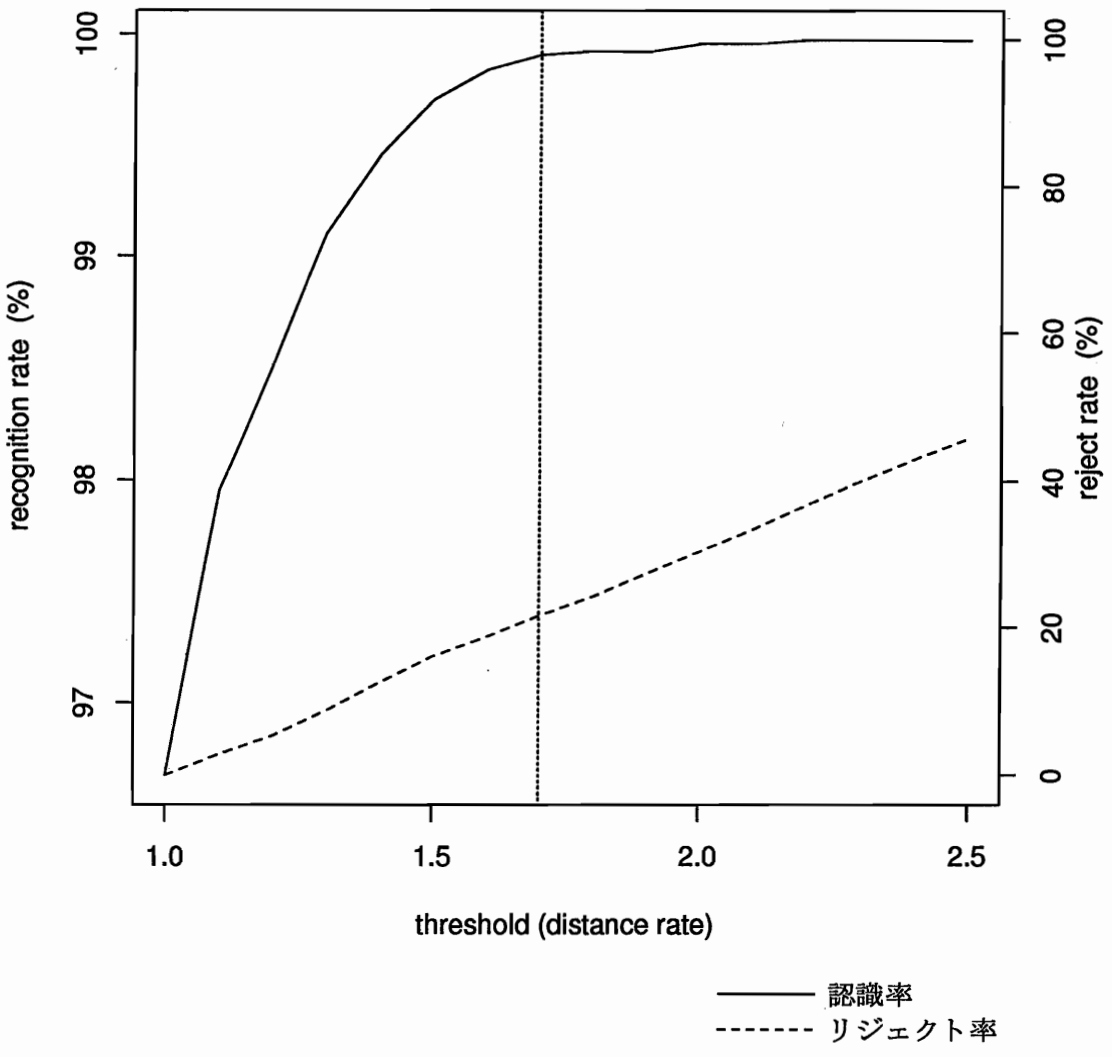

図 9: 1位、2位の距離比による認識率とリジェクト率

4.1.2 使用候補数

リジェクト文字に対して、第何位まで文字候補を考慮すればよいかを調べるため、後処理実験で使用する文書について、候補数と認識率の関係を調べた。(図 10) 辞書と文法により後処理を行なうため、記号 (句読点や括弧等) については修正できないので、それらは後処理対象から除外して考える。記号を除いた場合が図の実線である。

候補数を多く取る事により、正解文字を多く拾える事になると思われるが、考慮する文字が増えると、不正解の文字の組み合わせにより思わぬ単語、文節が出来てしまう事にもなりかねない。

グラフを見ると、候補数をあまり増やさなくても高い認識率となっているので、今回の後処理実験では、候補数として第 3 位までを考慮する。

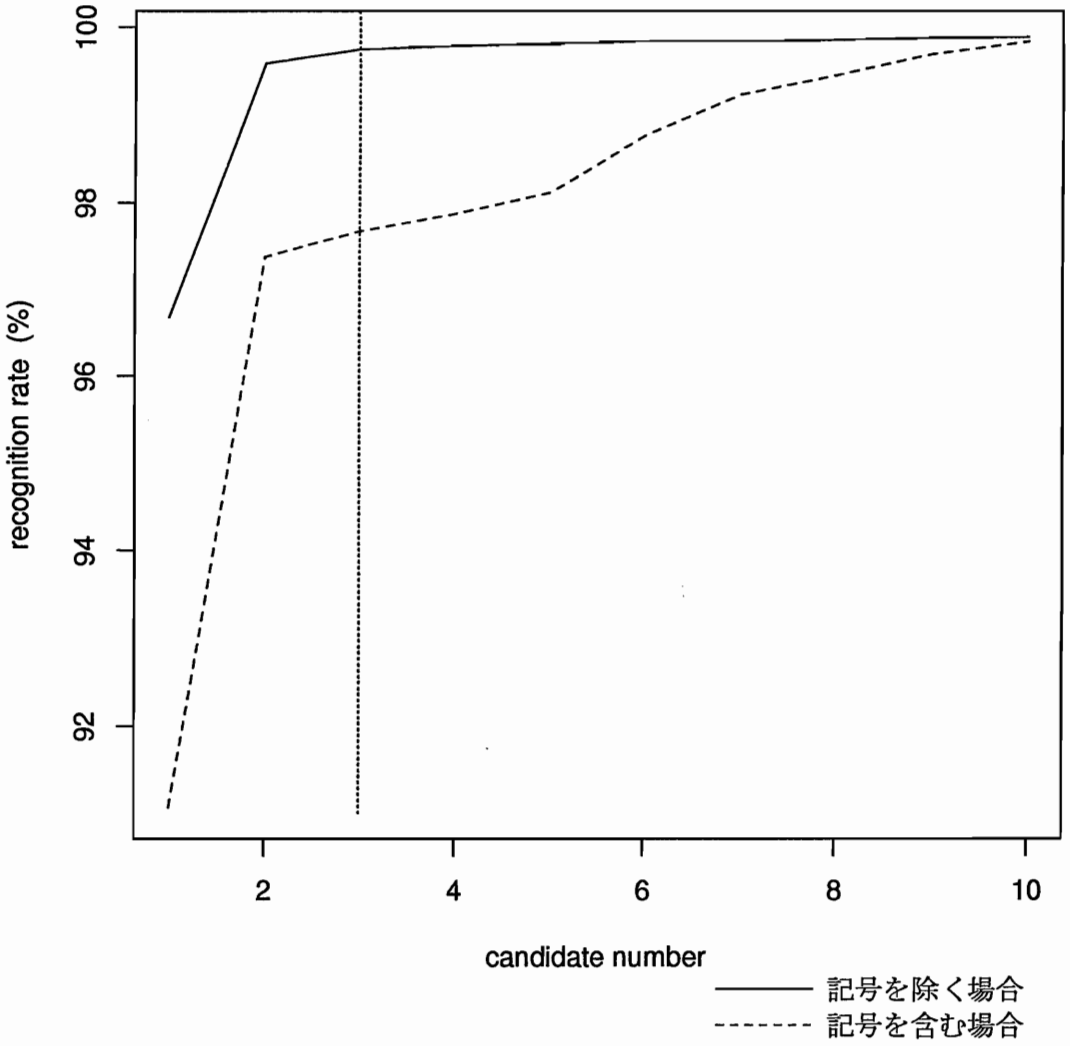

図 10: 候補数による累積認識率

4.2 後処理実験

4.2.1 データ入力

実験データとして、新聞の社説７５００文字(７５０文字×１０種類の文章。そのうち記号をのぞくと６９４９文字) を用意した。これを文字の大きさ１０ pt でレーザープリンターより出力したものをスキャナから読み込ませた。今回切り出しミスに対する対処はしていないため。文書の各文字を格子状の枠に入るように並べ、切り出しミスが起らない状態にした。(図 11) この文書の認識結果 (候補文字、順位、距離値) を入力とし、後処理を行なった。

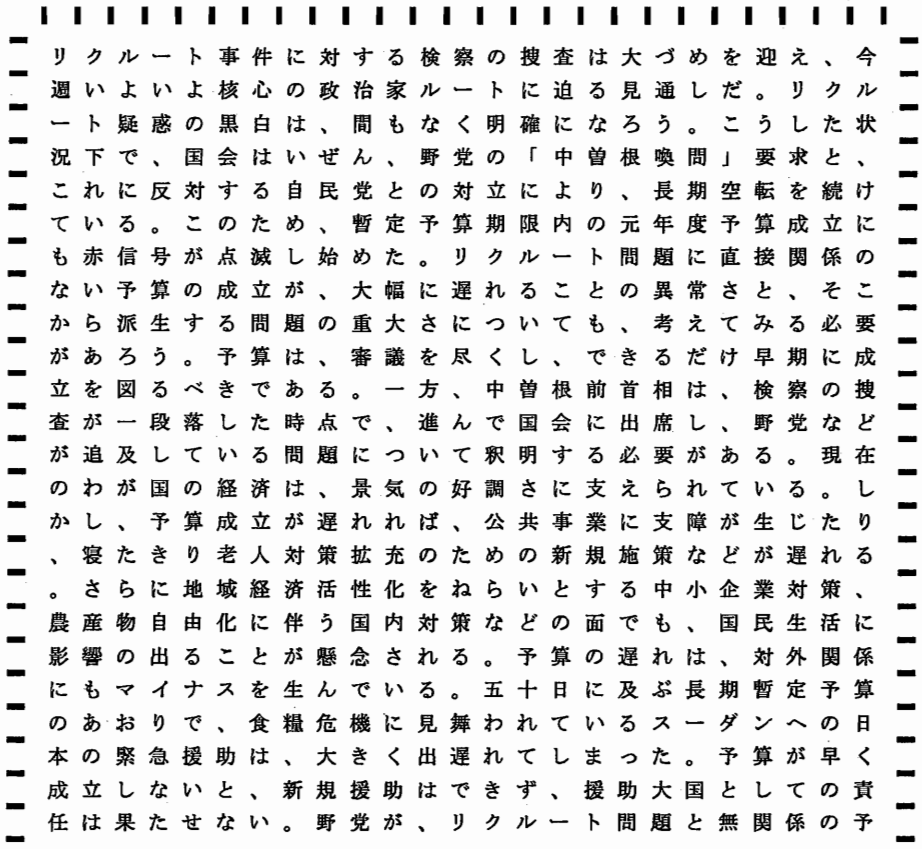

図 11: 読み込ませたデータ (例)

4.2.2 実験結果

閾値１．７によりリジェクトされた文字は、次の通りであった。

表 1: リジェクト文字の内訳

漢字	５２４文字
ひらがな	８１５文字
カタカナ	１２３文字
全体	**１４６２文字**

(全文字数の２１．０３%)

このうち、認識部で誤認識されていたが、後処理においてリジェクトされなかった文字が７文字存在した。(て、た、ト、一)

後処理による認識修正結果を文字種別に分類すると、次のようになる。

表 2: 後処理結果

	文字種	正	誤
文節として認識したもの	漢字	466	5
	ひらがな	361	3
	カタカナ	76	6
小計		903	14
文節としては認識できず１位候補を出力したもの	漢字	34	5
	ひらがな	302	54
	カタカナ	27	14
複数の候補があがったため１位候補を出力したもの	漢字	12	2
	ひらがな	15	80
小計		390	155
合計		1293	169

(単位：文字)

得られた結果をまとめると、次のようになる。

まず、リジェクトされた文字数に対する、正解文字数の割合を出した。この正解数の中には、文節候補が存在しないか、又は複数候補があがったため、文節として確定できず、1位候補を出力した際に、結果として正解になったものも含まれる。この割合をここでは正解率と呼ぶ。

表 3: 正解率 (正解数／リジェクト数)

漢字	９７．７０％
ひらがな	８３．１９％
カタカナ	８３．７３％
全体	**８８．４４％**

この割合には、単語照合はうまくいかなかったが1位候補がたまたま正しいものであったため、正解となったものもありうるわけである。そこで、後処理の効果を知るために、それらを除いた、つまり文節として認識したうえで正解文字を出力できた割合を次に求めた。これを確定率と呼ぶ事にする。

表 4: 確定率 (文節として確定できた正解数／リジェクト数)

漢字	８８．９３％
ひらがな	４４．２９％
カタカナ	６１．７８％
全体	**６１．７６％**

正解率においてもそうであるが、確定率においては特に、漢字に比べひらがなが低い数字を示している。これは、まずひとつには、自立語辞書に登録されている単語が、漢字中心である事が挙げられる。辞書に漢字で登録されている単語を、ひらがなで表記している場合、単語照合がうまくいかなくなってしまう。また、自立語が未登録である場合、その後に接続している付属語がすべて救えなくなってしまう、ひらがなが続く場合文節の切り出しが正確に出来ない、などの理由から、ひらがなに対する精度が低くなっているものと思われる。

誤認識の具体例や、その原因などについては、次章において深くふれる。

文節として認識できたとしても、それが正解であるとは限らない。誤った文字によって単語、文節が出来てしまう事も充分考えられるわけである。そこで次に、文節として確定した際に、それがどの程度正解であるのか、その割合を求めた。これを、ここでは確定文字正解数と呼ぶ。

表 5: 確定文字正解数 (文節として確定できた正解数／リジェクト数)

漢字	９８．９３％
ひらがな	９９．１７％
カタカナ	９２．６８％
全体	**９８．４７％**

確定率の低かったひらがなでも、一旦文節であると認識した場合、誤認識は少ないといえる。しかし、誤認識した例は数の上では少なかったが、これは認識部で正しく認識したものを後処理によって改悪してしまう事も起り得るわけで、あってはならない事である。

最終的な認識率を求めてると、

後処理前		後処理後
９６．６７％	→	**９７．４６％**

であった。

認識率の変化をグラフにしたものが図 12である。横軸が後処理前、縦軸が後処理後の認識率であり、後処理の効果がない場合には図の点線 (原点を通り傾き 1 の直線) 上にあり、効果があるほど点線より上に点がくる事になる。10 種類の文書をそれぞれについて表したものが △、すべて総合したものが◆ である。

それぞれに、認識率を向上させる事ができた事がわかる。

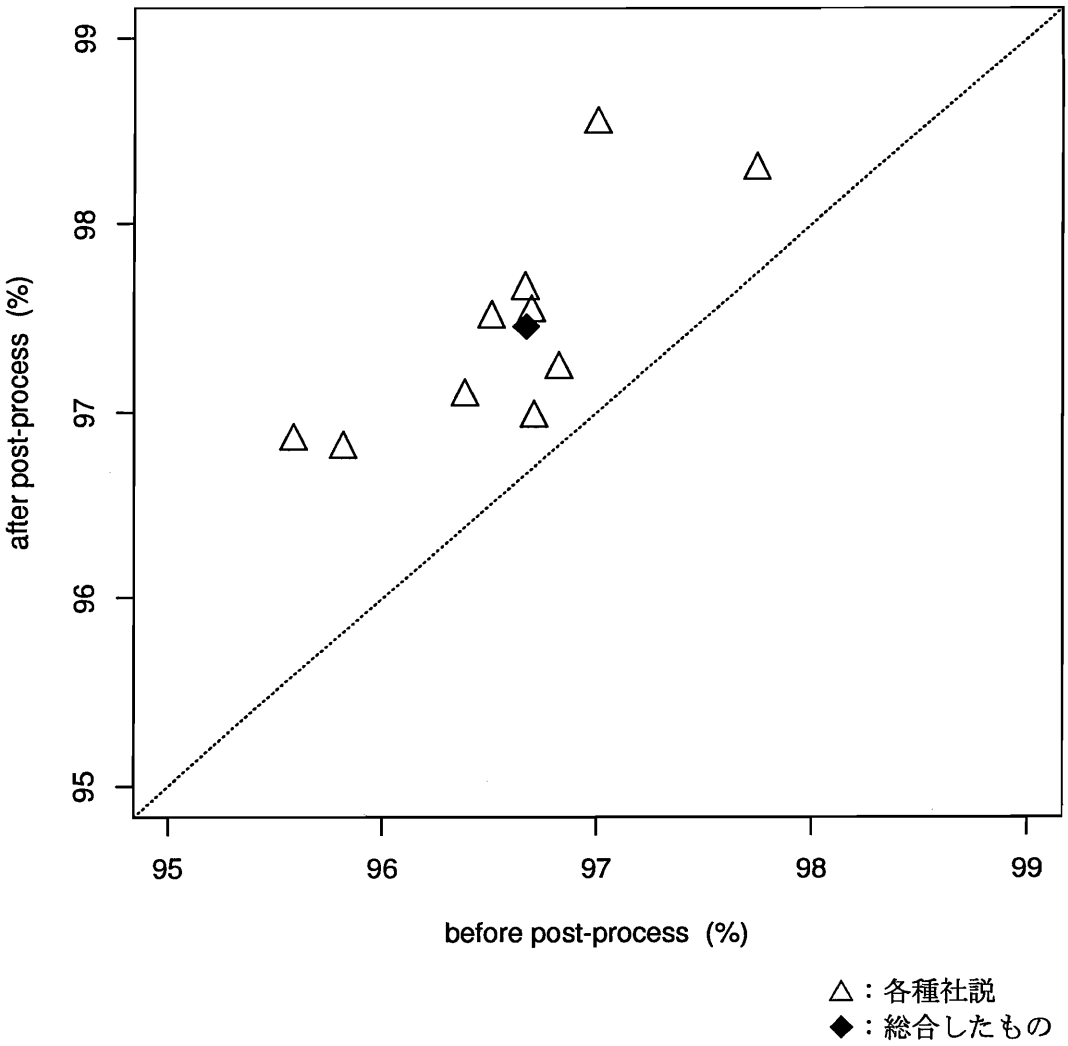

図 12: 後処理による認識率の改善

5　後処理不能文字の分析

今回の実験では、個々の文字についての間違えやすさや、文法の頻度などについては、考慮されていない。また、認識部からの認識情報も、充分には活かされていない。今後、後処理の研究をすすめていくにあたって、どのような点に着目すれば良いのか、今回の具体例を見ながら考察する。

5.1　改善した例

まず、今回の実験でどのような誤認識文字を救済できたか、その例を挙げてみる。

スキヤンダル　→　スキャンダル
労働時間　→　労働時間

などの、名詞や、

中国ては　→　中国では
取り組んている　→　取り組んでいる
積もつて　→　積って

など名詞の後の助詞や語尾変化などを救済している例が見られた。特に今挙げた「つ」と「っ」や、「て」と「で」などは、認識部で間違えやすい文字であるが、文法の規則より修正が比較的出来やすい例である。立{つ / っ}などの場合、自立語及びその語尾変化からでは、終止形及び連体形の「立つ」と連用形促音便の「立っ」が考えられ、確定できなかったが、「立っ」は'後ろに接続助詞である「て」が接続して文節になる'という規則を用いる事により、「っ」と判別できる。文法規則の効果がみられる例である。

自立語がうまく照合できれば、「立っていることを」「認めているように」「要求していたものだからだ」等、助詞、助動詞、形式名詞が二重、三重に接続したものも、文節として認識できる。基本的な文法には、強いといえよう。

5.2　改悪した例

認識部では正しく認識していたのに、後処理で誤った文字を出力した例に、次のようなものがある。

筑波大は　→　筑波犬は
ゴールデンウイークの　→　ゴールデンウイータの

これらは何故このように改悪してしまうかというと、「筑波大」「ゴールデンウイーク」という単語が辞書に登録されておらず、下位候補である「犬」「夕 (夕方の'ゆう')」という１文字の名詞がそれぞれ係助詞「は」、格助詞「の」と接続し、文節となってしまったのである。１文字名詞を他の名詞と同じ様に扱うとこのような事が起り得る。１文字名詞の出現頻度はあまり高くはないので、その取り扱い方に工夫が必要であると思われる。

一九七三　→　一丸七三

これは、「一」「九」「七」「三」は当然、それぞれ１文字ずつでしか登録されておらず、「一丸」という名詞が最長の文節として出力されてしまった例である。今回はもともと縦書きの文章のため、漢数字が使用されている。算用数字においても「数字のあとには数字がきやすい」「数字＋数字→数字」などの規則を用いることにより、防げるものと思われる。これは、カタカナや、アルファベットについてもいえる。

また、珍しい例ではあるが、次のようなものがあった。

あいまいに{ し / じ / ビ }{ か / が / ル }ち　(正解：「あいまいにしがち」)

これは「～にしがち」といった表現が認識できず、３位候補同士が接続した「ビル」という単語を認識してしまったものである。３位候補同士が結び付くということは滅多にある事ではないと考えられる。このような場合、再び距離値などの認識情報を考慮し、確からしさを確認する事が必要になる。

5.3　複数候補の出た例

感{ 覚 / 賞 }　{ 職 / 磯 }場　{ 資 / 賃 }金　国{ 連 / 運 }　減ら{ ず / す }　一{ つ / で }は

これらは異なった候補文字の組合わせのどちらからも文節候補が生成されてしまった例である。この場合、再度距離値の比をとり候補をしぼるといった方法も考えられるが、１位と２位が残っている場合にはこれではうまくいかない。仮に、文脈の意味を調べ、より存在しそうな単語を判定できたとしても、「資金」と「賃金」、「国連」と「国運」などは、同一文章内に同時に存在する事もあると考えられ、これらを判別するのは困難である。

そして、非常に目立った例として次のものがある。

結果{ か / が }

これは、名詞の後に格助詞の「か」及び「が」が接続し、文節となった例である。名詞の後の「が」という助詞は文章中に数多く現われるが、そのほとんどが認識部において１位候補が「か」と誤認識され、後処理でも修正できないという事態になる。このような例が実に７５例存在した。ひらがなの正解率、確定率をさげている大きな原因となっている。認識部で誤認識しやすいものであるため、後処理の効果を期待したい文字であるが、文法の構造上は似ているため、判別は難しい。出現頻度からいえば、「が」のほうが圧倒的に高いので、そのあたりが手がかりになるであろうか。また、接続して構成された文節が主語になるかならないかの違いなども手がかりとして考えられるものである。

5.4 文節切り出しミスの例

文節を切り出す段階で、ミスが起る事がある。次のようなものがあった。

リクルい／ト(リクルート)
リクルート／社会／長
課税制／度
十時／間

最初の例は、「リクルート」の「ー」が「い」と誤認識され、文節の切り出しは平仮名→非平仮名の部分で行うので、'い→ト'の部分で分離してしまったものである。切り出しを、1位候補の文字種でのみ行っているため、このような事態が起き、「リクルート」が辞書に登録されているのにもかかわらず修正できない事になる。

また、その次からの例では、例えば「課税制度」であったら、'課税／制度'が本来の単語の切れ目であるが、文の前から調べていくため、「課税制」という単語をまず認識して、残りの「度」をその後調べる事になる。「度」が1位候補に正しく認識されていない場合、「制度」として修正する事が出来なくなってしまう。

また、この文節切り出しの問題点として、平仮名→非平仮名の部分を含む文節の存在が考えられる。「組み立てる」「走り去る」などは、「組み＋立てる」「走り＋去る」として認識できるが、「開けっ放し」といった特殊なものもある。「お楽しみ」「ご自身」など、「御」をひらがなで書くという場合も多くみられる。接続しやすい単語にはその点を考慮する規則を与えるといった事が必要になるであろうか。照合範囲についても再考する余地がある。これに関連して、「～的」「～型」など、様々な単語に接続する文字や、「～時」「～個」「～割」など、数字の後に接続する文字について、それなりの規則を与えるという方法も考えられる。しかし、あまりその様な規則を増やしすぎると、誤認識された文字により文節が構成されるケースが増えてしまう。必要な規則を選択する事が要求される。

5.5 その他の例

今回の実験に用いた文章は、新聞の社説であり、比較的漢字の多く使われている文章であるが、それでも本来漢字で書くところを平仮名で表記するものがみうけられる。文章中にみられた「さまざまな」「やめる」「あらためて」等は、平仮名で表記するほうが多いと考えられる。自立語辞書ＫＩＤ－Ｊ８２には、単語の読みの情報も含まれているので、うまく使用する事によって、そのような例を修正する事ができるのではないだろうか。

また、実験をすすめていく過程で、最初は自立語と助詞、助動詞との接続だけを考えていたが、「こと」「ため」「とき」「まま」「よう」などの、形式名詞との接続も考慮する事によって、認識率の向上がみられた。だが、まだすべてが網羅されてはいない。「～という」「～にとって」「～において」などの表現は、これらを構成する単語本来の意味がうすれており、文法にこだわって単語を一語一語調べるよりも、ひとまとまりの表現として考えたほうが取り扱いが容易になる。今挙げたものは出現頻度も高く、効率よく精度をあげる事ができると思われる。

6　結論

6.1　結論

文字認識により得られる認識結果を修正するための情報となる単語、文法など言語情報の形式を決定し、またそれらを照合する手法を考案した。

この手法により、文字認識率をある程度向上させる事に成功した。リジェクトされる文字はひらがなが多いが、個々の単語の照合に加え文法規則を導入した事が、効果をあげたと思われる。しかし、ひらがなについては確定率が低く、まだ後処理前の認識に頼っている部分が大きい。今後、さらに手法の改善、言語情報の充実により精度を高めていく必要がある。

6.2　今後の課題

今後、研究をすすめるうえで、課題となる点について次に示す。

・単語辞書の充実

本実験で辞書をつくる基となった自立語辞書ＫＩＤ－Ｊ８２が作成されたのは、昭和５９年と古く、最近になり使われるようになった言葉が登録されていない (今回使用した文章では「フロン」「財テク」など)。また、固有名詞がほとんど無いといった欠点がある。未登録語は後処理の精度を落とす大きな原因となるため、改善すべき点である。

・文節の確からしさ

本実験の手法では、文節と思われる文字列を探すだけで、果してそれがどの程度確からしいものであるのかといった問題については考慮されていない。理想を言えば、'この文字列は未登録でわかりませんが、この文字列はこのような文章です。'といった事が間違いなく言える事である。これには、認識して得られる文字毎の評価値の有効利用などが考えられるが、文節の確からしさというものを、どのように評価するかというと、これは難しい問題である。

・効率の良いアルゴリズム

今回はすべての文字について、考慮する候補数、リジェクトの閾値などを一律にしたが、より後処理の効果を上げるためには、文字毎に間違えやすさや間違えやすい文字などを考慮する事が考えられる。もともとの認識率が高い事を前提とした場合、あまり多くのパターンを考え誤認識してしまうよりも、起りやすい間違いに的を絞った対処が効率を高めるためにも有効となるであろう。もともとの認識率が低い場合に同じ手法でよいのかどうかも、検討すべき点である。

・ 謝辞

本研究を進めるにあたり、全般的な御指導を賜わりました東北大学工学部阿曽弘具教授に心から感謝いたします。

そして、文字認識の専門分野におきまして御助言を賜わりました、東北大学工学部丸岡研究室 (阿曽グループ) の大町真一郎氏、後藤英昭氏、池田啓明氏、越後和徳氏、に深く感謝いたします。

さらに、日頃から本研究につきまして御討議、御協力いただきました同研究室の皆様に深くお礼申し上げます。

また、本研究におきまして『九州芸工大自立語辞書ＫＩＤ－Ｊ８２』を使用させていただきました。これにより、研究を大きく進める事ができました。改めまして、お礼申し上げます。

・参考文献

1. 長井 茂：「文字連接情報を活用する手書き文字認識の研究」
東北大学大学院工学研究科修士学位論文 昭和 62 年

2. 新谷、目黒、梅田：「認識情報及び単語・文節情報を利用した文字認識後処理
電気通信学会論文誌 '84/11 Vol.J67-D No.11

3. 首藤、楢原、吉田：「日本語の機械処理のための文節構造モデル」
電気通信学会論文誌 '79/12 Vol.J62-D No.12

4. 高尾、西野：「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 Nov.1989 Vol.30 No.11

5. 阪倉 篤義：「改稿 日本文法の話」 教育出版