自然言語処理 62－10
（１９８７・７・２４）

文字認識後処理の可能性

西野 文人
（富士通研究所）

日本語文書の文字認識の後処理を人間が行なった場合にどのくらい認識率が向上するかを実験した．従来の機械による後処理では，単語の照合検査と単語間の接続情報を用いることによって，後処理前の認識率が９０％程度の場合には９９％程度の認識率へ向上するが，後処理前の認識率が８０％を下回るような低認識率のデータに対しては後処理をしても１％程度の認識率の向上しか得られなかった．そこで今後の文字認識後処理の可能性を探るために，低認識率のデータに対して，人間による後処理を実験したところ，９３～９９％程度の認識率を得ることができた．また文脈がわからなくなるように細切れにした文章でも８５～９０％程度の認識率を得た．本稿ではこれらの実験の結果を述べるとともに，今後の文字認識後処理の課題と可能性について考察する．

Postprocessing Potential for Character Recognition

Fumihito NISHINO

Software Laboratory, Fujitsu Laboratories Ltd. Kawasaki
1015, Kamikodanaka, Nakahara-ku, Kawasaki 211, Japan

This paper discusses experiments in human postprocessing of Japanese text for character recognition. If the recognition rate is 90%, it is possible to improve the rate to 99% with word-connection information. However, if the recognition rate is less than 80%, only a 1% improvement was attained. If this same data is subject to human postprocessing, a 93% to 99% recognition rate is attainable. This paper shows these experiments and discusses the problems.

1. はじめに

日本語文書の文字認識の後処理では，文字認識で得られた結果をもとに言語的処理を行ない，文として正しくつながる文字列を求めることによって正しい候補を選択する．この言語的処理として，現在のところ実行時間や技術的な面から，主に単語辞書との照合検査と単語間の接続関係の検査によるもの［1～6］や2文字連接確率を利用するもの［7］といったレベルのものが使われている．

文字の認識率が90% を超える場合には，この程度の後処理でも99% 程度まで認識率を向上させることができる．しかし文字認識率が 80%を下回り文字認識の候補中に正解文字が存在しないこともあるような低認識率のデータに対する後処理では，単語の照合検査と単語間の接続レベルのものだけではわずか1%程度しか認識率を向上させることができなかった．

きれいな印刷文字に対する文字認識では後処理前の認識率が90% 以上を実現することは容易であろう．しかし，コピー等によるにじみ，かすれ等の低品質文字に対する文字認識や手書き文字の認識では，文字認識率が90% に至らないこともあり，このような文字認識率が低い場合のことも考えておく必要がある．

このような文字認識率が低いデータに対して，人間が後処理をしたら正解の文字を推測できるだろうか．推測できるとしたらどのような情報を使っているのだろうか．文字認識そのものに関する研究として，OCRと人間の文字認識能力とを対比した報告はなされているが［8］，文字認識の後処理に対する人間の能力についての報告はない．そこで，

(1) 人間の後処理能力を知ることによって，文字認識の後処理の能力の可能性（どこまで認識率を高められるか，どの程度まで低い文字認識率であっても救済可能か）を知り，機械による後処理の目標とする．

(2) 人間の後処理の推論方法を探ることによって，後処理アルゴリズム開発の指針とする．

ことをねらいとして，人間による後処理実験を行なった．本稿ではこの実験について述べ，その結果をもとに今後の文字認識後処理の課題と可能性について考察する．

2. 人間による文字認識後処理の実験

文章に対する文字認識後処理の性能は，次の条件で変化すると考えられる．

文章の要因

・文章の読み易さ

・文体

文字認識の要因

・認識率

・候補文字数

・正解文字と不正解文字との区別

後処理プロセッサの能力

・文法知識

・一般常識・専門分野の知識

さらに，人間によるシミュレーションでは次のようなことでも性能が変化すると考えられる．

人の要因

・実験への取り組み態度

そこで，高卒以上の日本語を自由に使いこなせる人を被験者として，次のような条件を設定して実験を行なった．

① 低認識率データに対する人間の能力を調べるため，コピーを繰り返して文字を低品質化させた文書を読み取らせた認識率(a) 76.8%，(b) 80.9%の文字認識出力データ（659 文字，各々20位まで候補を出力．付録Aにデータの一部を示す）に対して人間による後処理実験をした．

② 文脈情報がどのくらい重要なのかを調べるため，①(a)のデータの一部を単語ないしは2～3の文節単位にランダムに切り貼りしたデータ（付録B）に対して人間による後処理実験をした．なお，このデータは不正解を含む部分だけを取り出したので部分的には認識率は55.8% である．

③ 候補文字の数の依存性を調べるために①(a)のデータを第1位の候補文字のみを出力したデータに対して人間による後処理実験をした．

④ 認識率の違いによる後処理性能の違いを調べるために，各種の認識率のデータを作成（①とは別のテ

キストの一部をランダムに□で置き換えたデータを機械的に作成）し，それぞれに対して人間による後処理実験をした．

3. 実験結果と考察

後処理の性能を比較するために，救済率［後処理前の第1位候補が正解でない文字数をａ，後処理後の正解でない文字数をｂとしたとき，（ａ－ｂ）／ａ］で示す．

3.1 実験結果

単語照合と接続検定による機械処理では，認識率90.6% のデータに対しては救済率90.3% （後処理後認識率99.1% ）が得られたが，実験①(a)の認識率 76.8 % のデータに対してはわずかに6.5%の救済率（後処理後認識率78.3% ）しか得られなかった．

これに対して実験①に対する救済率は，(a)認識率76.8% のデータに対しては92.6，81.5，74.8，69.0，67.4，63.7，63.7% (b)認識率80.9% のデータに対しては99.4，87.3，83.3，69.0% であり，機械による後処理に比べて非常に高い救済率を出せることがわかった．（図1）

人間の後処理では，「構内⇨社内」，「速い⇨遅い」のように意味的に似ているものと間違えることはあったが，機械による処理ではほとんどお手上げの状態であった「どの文字が正しく認識されていて，どの文字が誤って認識されているのか」の判断はほとんど間違えることはなかった．その結果，人間による単語の切り出しは正確であり，たとえ部分的に正解の単語がわからなくても，不明な部分が他の正解部分を誤った結果に導いてしまうことはなかった．

一方，文脈を取り除いた②に対する救済率は，83.9，60.9，50.6，48.3，35.6% であり，これでも機械による後処理に比べれば救済率が高いことがわかった．しかし，実験①に比べると救済率にして25% 程度低いものになり，文脈処理が重要な要素であることがわかる．

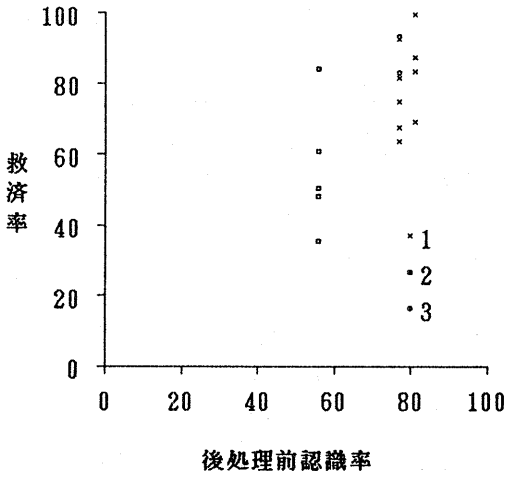


図1 人間による後処理実験の結果

候補文字数を第1位のみ出力した③に対する救済率は，93.3，83.0% であり，実験①(a)の結果と変りなく，人間による後処理においては候補文字数はあまり重要な要素でないことがわかる．付録Ｃに93.3% の救済率が得られた人間による後処理結果を，表1にこの被験者が後処理で救済できなかった部分を示す．

認識結果	人間の後処理	正解
む肉交換浴	社内交換機	構内交換機
打８操作	打？操作	打鍵操作
８括絵 ケーブル	電話機 ケーブル	電話線 ケーブル
品労保偉	品質保持	品質保障
住友寓工 株式会社	住友重工 株式会社	住友電工 株式会社
北俣営姓 妹式会社	北俣電機 株式会社	北陽電機 株式会社
品矢保陰	品質保証	品質保障
変位ｉを	変位ｉを	変位量を
走査方式は… 連傍酌な枚査	連続的な走査	連続的な検査

表1 救済できなかったもの

次に，実験④の結果を表2及び横軸に後処理前の認識率，縦軸に救済率をとってプロットしたものを図2に示す．個人差はあるものの，このデータから認識率

n%のデータに対して少なくとも(n-20)% 以上の救済率を得ていることがわかる．ここで，表2において高い救済率を出した被験者e，iは新入社員であり，実験に対する取り組み態度の違いといえるかもしれない．（被験者iは付録Cの後処理結果を与えた被験者である．）

被験者	後処理前認識率	後処理後認識率	救済率
a	99.2%	99.9%	85.7%
b	95.2%	98.3%	65.9%
c	88.1%	96.1%	67.6%
d	83.5%	94.4%	66.0%
e	80.9%	96.7%	82.8%
f	72.2%	88.1%	57.1%
g	69.3%	85.5%	52.7%
h	62.6%	81.7%	51.2%
i	56.3%	91.9%	81.4%
j	53.4%	73.9%	44.1%
k	45.8%	57.4%	21.5%
l	37.5%	52.6%	24.1%

表2　後処理前認識率と救済率の関係

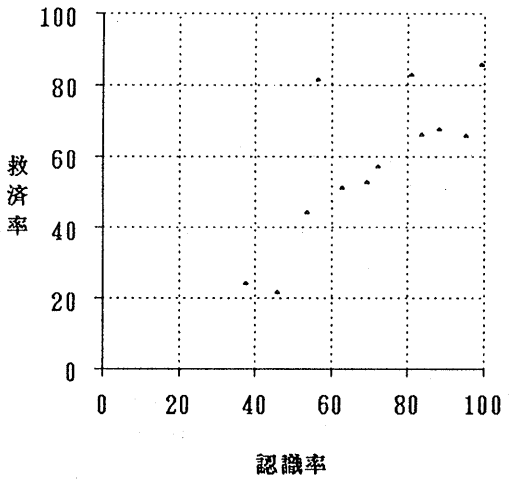


図2　認識率と救済率の関係

なお，実験④では実験データ作成上の都合で，

(a)　正解文字と不正解文字を明確にした．

(b)　不正解文字に対する候補文字を与えない．

(c)　実験①～③とは異なる文書を使った．

(d)　不正解文字はランダムに決定した．

という点で実験①～③とは異なっているが，これらの違いは救済率に影響を与えているだろうか．そこで，実験①のデータを上記(a)(b)の条件にして追加実験⑤を行なったところ67.4% の救済率が得られた．実験④⑤から(c)(d)の違いは文法能力の十分ある人間にとってはあまり影響がないと言えよう．(a)の違いは機械にとっては後処理能力を向上させる大きな要素であるが，人間は文字認識結果に対しての正解か不正解かの判断にはほとんど間違いがないので後処理能力に大きな差は与えていない（ただし，たまたま意味が通じるような誤った文字が第1位候補として出力された時にその誤りを見落す［表1の「変位i」，実験⑤の被験者は「変位量」という正解を出した］）．(b)は，人間では多くの場合は出力結果の候補文字の形に関係なく正解文字を推測しているが，推測される候補が複数あった時や候補の推測ができなかった時には，候補文字の形を利用したり他で同様な誤りをしているところを探すなどして候補文字の情報を利用しているので，候補文字を与えないことによって若干救済率を下げているであろう（特に認識率の低いデータに対して）．

3.2　人間の推論方法

以上の実験結果が示すように，人間による後処理ではかなりの救済率をあげることができることがわかった．この節では人間が後処理を行なう際に利用している情報について考察し，機械で同等の処理をする際の問題点について考察する．

(1)　文章構成上の重要な特徴の利用

機械による後処理と人間による後処理との大きな性能の違いは，機械では一部で全くわからない部分があるとそこを何とか辻褄を合せようとして誤りを周辺に伝播させてしまうのに対して，人間では助詞や助動詞等の文章を構成する上で重要な要素となるものをしっかり認識し，文章の構造をはっきり把握しているので，たとえ元の単語がまったくわからないような状況でも単語分割や構文誤りを犯すことはないという点である．

機械で後処理をするのにも，単純に前から処理するのではなく，構文決定上重要な単語と非重要な単語とを区別して，重要な単語を中心に構造決定をまず行な

うようにするのが有効であろう．

(2) 構文情報の利用

人間の場合，短い文章である限り品詞や係り受けの正当でないものを見つけることができる．

例) 付着をどの表面欠陥を検査することが・・

機械による後処理でも構文解析を行なうことは必要である．ただし，一般に行なわれている構文解析は正しい入力文に対する解析であるのに対して，日本語文字認識後処理では入力文の誤りを見つけるための解析であるという点と，標準日本語文法にのっとっていないデータ（新聞見出し，詩歌等）に対しても読み取る必要があるという点で，今までの解析技術と若干異なる技術が必要である．

(3) 単語情報の利用

人間の場合，正解文字が候補文字にない場合でも，適当な数文字をキーとして，適当に思い付く単語を探している．

例： かな○字変換 ⇨ かな漢字変換

このため，実験③のデータが示すように人間にとっては候補文字の数はあまり関係がなかった．しかし機械で同等のことを実現するためには，認識不能単語を抽出し，誤読の可能性の高い文字を除いた部分で部分照合による辞書検索を行ない，単語を確定する必要がある．そのためには，「○話番号」，「○気○号」などの文字列の一部が不明であるようなものをキーとする高速な辞書検索手段が必要である．この処理を短い単語に対して行なうと意味検査等を行なわなければ誤った後処理をしてしまう可能性があるが，複合語や専門用語のような比較的長い単語には十分有用であろう．また，推測した文字が2通り以上あるような場合には，元の文字認識部に対して，どちらがより正しそうであるかの検査を依頼するような機構も必要である．

(4) 語と語の関係情報の利用

人間による後処理の場合には，単に単語の情報のみではなく，単語と単語の結びつきの強さを利用していることが多い．

例：「手間が大俣に短柿する」

「手間が」という言葉に対して後続する単語を考えると「かかる，へる，はぶける，多い，…」というようにかなり限られている．このような語と語の関係を集めたデータ［9］は，文字認識後処理には有効であろう．

(5) 関連語情報の利用

人間の場合，文章全体での話題を的確に把握できる．これは単語の断片だけを与えた実験②においても全体の話題を把握することができる．これにより，最初は全くわからないような単語も，後ろの方を読んだ後で推論することが出来る．

例：「シグマホンはマイコン内蔵型＊＊機です．」

⇨ 「シグマホン」の「ホン(PHONE)」から電話の話題であることを認識し，「＊＊」が「電話」であることを認識する．

これを実現するためにはシソーラス等を利用して関連語（電話⇨受話器，電話機，電話線，交換機，・・・）を検索して適切な単語を選択することが必要であろう．

(6) 文脈処理の利用

指示詞情報も後処理には有効である．

例：・～の電話機です．この＊＊＊は～

・欠肪を～．従来，これらの欠陥を～

「この＜名詞句＞」，「その＜名詞句＞」といった参照表現が使用されている場合，これらの＜名詞句＞は，その文章の前に現われていることが多い．このことを利用することによって，指示詞の直後の名詞句が認識できなかった場合には，その指示詞が現われる前の文章中を探すことによって，指示詞の直後の名詞句を認識できることがある．

これらの名詞句の認識には，外部世界に関する知識とその推論を用いなくても，「参照表現が段落の先頭の文中にあるとき以外は被参照表現が段落の範囲外にでることは少ない」とか「主題化されている概念は他の概念より指示されやすい」といったヒューリスティックス［10］を利用することによって，表層上の情報

だけを利用して検査するのであれば容易に実用化できるであろう.

(7) 専門知識・その他の情報の利用

人間の後処理では推論のために上記の他に様々な知識を利用している.

例：・「○○○のシグマホン」⇨シグマホンを作っている会社を知っている. ないしは, 具体的にどこの会社が作っているかを知らなくても, 電話機を作っている会社を知っていて文字数や文字の形の情報から○○○を富士通と推論する.

・「国際単位系（○Ｉ）」⇨国際単位系がＳＩと呼ばれることを知っているか, système international d'unitésという単語を知っていれば後処理可能である.

・「こう（○）配」⇨括弧の使い方としてその単語の漢字表記を示すことから○が「勾」であることが推論できる.

・「1000Kgf/s(9.80…500Kgf/s(○.○○…」⇨比例計算して後処理可能である.

・段下げや箇条書きなどの構成の情報や(1), (2)といった順序の情報等からも後処理が可能である.

5. まとめ

機械が救済不可能であった認識誤りでも, 人間が見れば修正可能なものも少なくない. 人間による後処理では80% の認識率に対して60～99% の救済率をあげることができた. また, 文脈を利用しなくても35～84% の救済率をあげることができた. そして60% の認識率であっても50%, 頑張れば80% の救済率をあげることができた.

人間が後処理をする場合と計算機でいかに処理するかということは必ずしも同一であるとは限らない. しかし, 人間がどのように判断, 処理しているかを知ることは, 今後どのようなことをしなければならないかという方向を見る上でも重要なことである. 今後の文字認識の開発に伴って, 文字認識後処理に与えられた課題は多い.

謝辞

本研究に対して貴重な助言及び実験に協力いただきました中西道明部長付, 内田裕士室長, 山本栄一郎主任研究員及び富士通研究所の方々に感謝いたします.

参考文献

[1] T. Kawada, S. Amano, K. Sakai: "Linguistic Error Correction of Japanese Sentence", COLING80, pp. 257-261 (1980)

[2] 新谷, 梅田：”文字認識における複合後処理法の能力評価”, 電子通信学会論文誌 Vol. J68-D, No. 5, pp. 1118-1124 (1985)

[3] 村瀬, 新谷, 梅田, 小高：”言語情報を利用した手書き文字列からの切出しと認識”, 電子通信学会論文誌 Vol. J69-D, No. 9, pp. 1292-1301 (1986)

[4] 水上, 岡田, 小林, 南部：”単語の文法的接続情報を利用した日本文認識の後処理”, 電子通信学会総合全国大会 No. 1547 (1986)

[5] 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, pp862-869 (1985)

[6] 伊藤, 加藤, 高橋：”文字認識における簡易後処理方式”, 情報処理学会第34回全国大会 pp. 1849-1850 (1987)

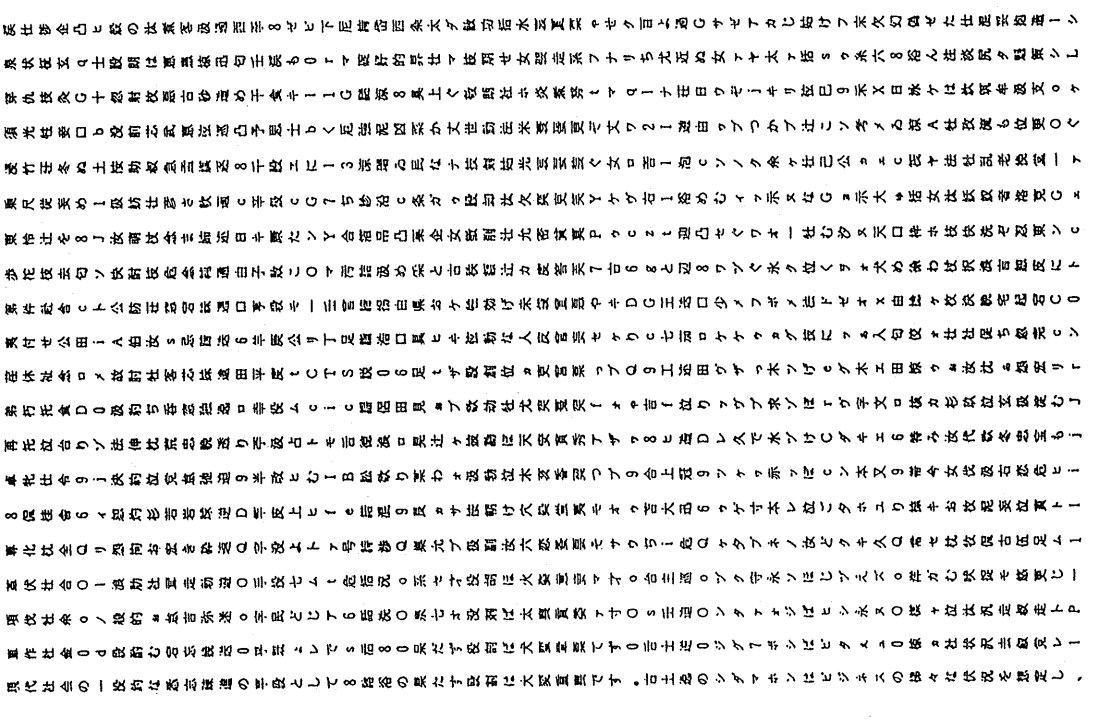
[7] 杉村, 斎藤：”文字連接情報を用いた読取り不能文字の判定処理── 文字認識への応用──”, 電子通信学会論文誌 Vol. J86-D, No. 1, pp. 64-71

[8] 飯田, 小森：”人間の文字認識能力の評価－手書き片仮名, 英数字に対する認識能力について－”, 電子通信学会論文誌 Vol. J67-D, No. 3, pp. 257-264 (1984)

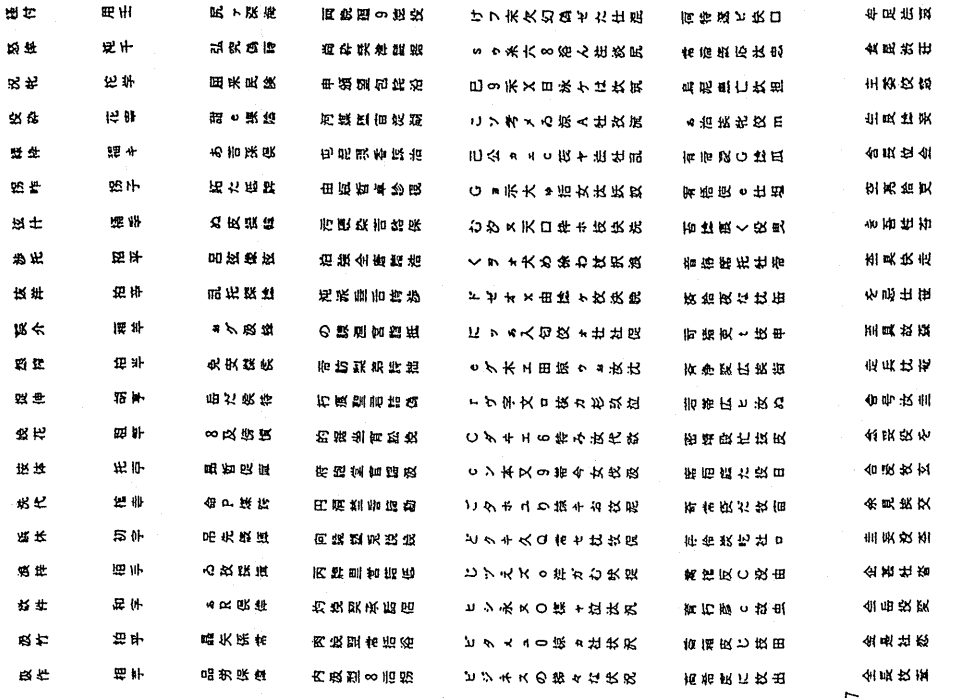
[9] 田中, 吉田：”知識データ（語と語の関係）による多義性の解消”, 情報処理学会自然言語処理研究会60-3 (1987)

[10] 永田, 辻井, 長尾：”日本語論説文に現われる照応表現の処理”, 情報処理学会34回全国大会 pp. 1225-1226 (1987)

付録A　実験1に用いた低認識率データの一部



付録B　実験2に用いた文脈を取り除いたデータの一部



付録C　実験3の後処理結果例