ヒューマンインタフェース 38－5
（1991. 9. 12）

OCR入力された日本語文の誤り検出と自動訂正

伊東 伸泰, 丸山 宏
日本アイ・ビー・エム(株) 東京基礎研究所

著者らは既存の日本語印刷文書をハイパーメディアなどのデータベースに効率よく入力・運用する目的で文書理解システム (DRS:Document Recognition System) を開発中であるが、その機能の1つとしてOCRで認識された文字中から日本語文の制約を利用して誤りを検出し、より確からしい候補に置き換える後処理を実現した。本後処理は日本語辞書と品詞間接続テーブルを参照して文法的に成立する文字列の候補を生成した後、各単語の品詞、出現頻度、遷移確率、および認識の確からしさに基づいてコストを計算しその値が最良のものから一定値以内の候補パスを選び出す。そして各カラムの文字候補について、自分自身を通る候補パスに付髄するコストと他の候補を通るそれから確信度を計算し、その値により当該候補の入れ替えや、オペレーターに対する警告を行う。実験によれば後処理なしで95%程度の認識率であったデータで認識率が約99%に向上し、検出されなかった(言い換えれば入れ替え、警告のいずれも行なわれなかった)誤認識文字は0.2%程度にとどまった。候補パスを見出す探索にはダイクストラ法とビームサーチを用いることで、80386(25MHz)のパーソナルコンピュータ上で約27文字/秒の実行速度が得られた。

A Method of Detecting and Correcting Errors in the Results of Japanese OCR

Nobuyasu ITOH, Hiroshi MARUYAMA
Tokyo Research Laboratory, IBM Japan Ltd.
5-19, Sanban-cho Chiyoda-ku Tokyo 102 Japan

Abstract

This paper deals with a post-processing method of the character recognition (particularly Japanese) developed as a part of Document Recognition Systsem (DRS), which is a efficient tool for entering printed documents into databases. The process consists of the following steps: (1) Recognition results are sent to the post-processor as a candidate set for each character. (2) Search the candidate lattice for existing words by consulting word dictionaries and the table of grammatical constraints on Japanese. (3) Make a search tree of possible paths. (4) Estimate a cost function of each path according to a part of speech, frequency (unigram and bigram) of each word, and reliability of each character in recognition process. (5) Select the candidate paths on the basis of total costs. (6) Calculate a certainty of each character from the two summations of the costs. One is that of the paths passing the character, and another is that of the paths passing other characters. (7) According to the certainty, replace the first candidate with more probable one, or give a warning to an operator. Our test shows that the recognition rate went up from 95% to about 99% and that only 0.2% of all characters were missed though they were misrecognized. The processing speed is 27 chars/sec on a personal computer (80386, 25MHz), which is acceptable to practical systems.

1 はじめに

日本語入力の手間を削減する方法としてOCRはきわめて有力である。しかしながら認識誤りを完全に避けることはほとんど不可能であり、入力後の確認・修正が不可欠である。したがって、入力文書が帳票ではなく1ページ当たり2000字程度は普通であるような一般雑誌等になると、OCRの認識速度よりもこの確認・修正作業の時間で全体の入力効率が左右されることになる[1]。そこでオペレータによるこの作業を補助および(半)自動化する試みが行なわれてきた。その中で比較的基本的なものは認識結果の確信度を識別時の距離等から算出し、結果が唯一に決められない場合はリジェクトとしてオペレータに警告すると同時に、その前後および周辺の文字から得られる制約をもとに候補文字の中から正解を推定するものである。利用する制約としては文字単位での連接情報[2]や単語としての成立可能性、さらに単語間の接続規則[3]などが報告されている。ところが入力文書の品質がよほどよい場合を除けば、きわめて多くのリジェクトが出力されてしまい、これらの手法が適用し難い場合も多い。そこで認識率が比較的低い場合にも適用可能な方法として、各候補文字を組合せてできるパスを日本語辞書と単語(言い換えれば品詞)間の接続規則を利用して探索する手法が提案された([4],[5])。この手法を適用するにあたって考慮しなければならない点としては、つぎのようなことが考えられる。

1.適用する日本語文法: 文字認識では対象となる文書を極端に絞ることは実用的ではないため、なるべく広い範囲の日本語文を受理できることが望ましい。ところが文字認識結果に対する制約として利用する場合には'ゆるい'文法であるほど、その効果が低下すると考えられる[4]。

2.処理速度: 現在のOCRの認識速度は10-100文字/秒程度であり、多くの場合そのエラー修正はパーソナルコンピューター上で行なわれるであろう。したがってパーソナルコンピューターで上記の速度に大きく遅れない程度の処理を認識と同期して行うことが要求される[5]。

3.得られたパスの評価: 候補文字の組合せから得られる(少なくとも文法的には正しい)パスは多くの場合複数存在する。そこで何らかの評価値(以下ではコストと呼ぶ)によって'より良い'パスを選択し、オペレータに提示する必要がある。さらに高尾ら[5]が将来の課題として述べているように後処理によって如何に認識率が向上するとしても100%になることはあり得ないのでオペレータによる確認は欠かせない。したがって後処理自身がその結果を評価し誤りらしい個所を指摘することができることが全体としての入力速度向上のために必要である。

筆者らは現在印刷文書を効率的にデータベース化するための文書理解システム(DRS: Document Recognition System)を開発中であるが、そのために必要な機能の1つとして、これらの要求を考慮するとともにDRSの目的に適した後処理を含む文字認識機能を実現したので報告する。最初に次節でDRSの概略を述べ、そのつぎに後処理の実現している機能および手法について説明する。さらに後処理の効果および速度についての実験結果を提示し、最後にまとめを行う。

2 DRS(文書理解システム)の概要

前述のようにDRSの目的は印刷文書(特に需要が大きいと思われる科学技術文献)をハイパーメディアなどのデータベースに効率よく入力することであり以下のような機能をもっている[6]。

1.レイアウト理解: 文書のレイアウト構造を与えられた文書モデルに基づいて解析し、書誌情報の抽出・読み順の決定を自動的に行う。これには図を自動的に検出しイメージとして取り出す機能も含まれる。

2.文字認識機能

3.認識誤りの検出・自動修正を行う後処理機能

4.キーワード候補の抽出: 文字認識すると同時に、文書検索に欠かすことのできないキーワードの候補を抽出する。

5.後処理と同時並列的に実行可能なエラー修正のためのユーザインターフェース

この中で2のみが漢字OCRアダプターカード (マイクロプロセッサ68020、3Mbyteのメモリーおよび専用ハードウェアからなる) 上で実行され、その他はすべてパーソナルコンピュータ (80386, 25MHz) 上でOS/2のもとに実現されている。文字認識単独の速度は約30文字/秒である。

3 後処理方式

3.1 日本語文法

池田ら [4] はOCRの後処理という立場から形態素レベルでの日本語文法を考察し、カテゴリー数86にのぼる品詞分類とその接続規則を提案している。しかしながら、'はじめに'で述べたように、より多くの文を受理することとより強い制約となることは相反する要求である。この事実と3番目の要求を考慮すれば、すべての接続規則を対等に扱うのではなく、文法自身に確率を付与することによってパス選択のときに利用するコストの1つとして取り入れることが必要である。機械翻訳の前処理としての形態素解析では、解が多くなり過ぎて次段である係り受け・構文解析に負担がかかり過ぎるのを防ぐため、単語間の接続に対して出現頻度や共起確率に基づいたコストを導入し、それぞれの解に付随するコストで解を序列化しようというコスト付き形態素解析の試みが報告されている (たとえば [7]) 。この場合でも単語をどのように分類するかは大きな問題となる。つまり分類がより細かい方が制約としてはより効果的であるが、信頼できる共起確率を求めるためにはBigramの場合でもカテゴリー数の二乗に比例して学習データ量を増やさなければならない。実用的な立場から言えばごく簡単な分類のBigramでも十分な制約になり得る場合もあれば、Trigramさらには複数文節間の関係を評価 (言い換えれば構文解析) しなければ妥当なコストを付けられない場合も存在するわけで、最も困難な場合にすべてを合わせることは実際的ではない。そこで原則は仮名漢字変換向けに開発された品詞分類 [8] を用い、誤認識されやすくかつその分類で同じカテゴリーに属している単語については必要に応じより詳細な分類および接続コスト (必要ならばTrigramや複数文節間にまたがる評価も含む) を記述できる枠組みを用意することにした。すなわち詳細分類のための辞書を別に用意しそこに記述されていなければ各品詞間の接続ごとに定義されているデフォールトのコストを用いることになる。それ以外の辞書は次のとおりである。

・**自立語辞書:**約115000語、自立語を39に分類

・**付属語辞書:**約900語、付属語を70に分類

・**ユーザ辞書:**現在は主としてコンピュータ関係の用語を格納

よく知られているように自立語(特に名詞)はその語数の多さの割には分類項目が少ない上、より詳細な分類が困難である。そこで共起確率は品詞ごとに計算するが出現頻度は各単語ごとに計算しその対数値に基づいたコストを、辞書の各エントリーに記述することにした。 学習に用いたのはJICST科学技術データベースの電気工学編(Vol. 26)である。各辞書はTRIE構造を採用しており、辞書引きを行う位置から前方の文字ラティスの要素のいずれかと適合するすべての長さの単語が高速に抽出できる。

3.2 パスの探索戦略とあいまい度の評価

最初にコスト付き形態素解析を記号を用いて形式的に表現し、次にその拡張としての後処理手法を示した後、最も重要なあいまい性の評価について述べる。 用いられる文字集合を$Cset$で表現すると単語(W)、文(S)はそれぞれ$W = p_1p_2,\ldots,p_l$, $S = q_1q_2,\ldots,q_m(p_i, q_j \in Cset)$と表現できるので、コスト付き形態素解析とは$S$を単語列として$S = W_1W_2,\ldots,W_n$のように分解し、合わせてその単語列から決まるコスト関数$g(W_1,\ldots,W_n)$を算出し、その値が最小コストから一定値以内であるか上位N位までに属するものを求める作業である。 ここでq_jを文字ではなく順序付けられた文字集合$Q_j = [q_{j1}, q_{j2},\ldots,q_{jK}]$([]は順序付けられていることを示すために用いる)に置き換えれば、文は文字列から文字ラティスとなる。通常の形態素解析では各文字位置(j)ごとにその先の部分文字列$q_jq_{j+1},\ldots,q_m$について辞書引きが行なわれるわけであるが、その替わりに部分ラティス$Q_jQ_{j+1},\ldots,Q_m$から得られる文字の組合せについて辞書引きを行ない候補単語を生成する手続きがあればその分場合の数は増加するが上記作業は容易に文字ラティスからコストの低い順にパスを求めるOCRの後処理手法に拡張できる。高尾らの後処理[5]はこの探索にA^*アルゴリズムを用いた最良のパス1個を見出すコスト付き形態素解析と考えることができる。 次に誤りらしい個所を指摘する機能を実現するため各候補文字のあいまい性について考える。図1(a)の文字ラティスを見れば(誰でも直感的に)'文書について'が正しい元の文ではないかと考えるであろう。しかしながら'文害'という言葉も'文(名詞)'+'害'(名詞)と考えれば文法規則を満たしている。さらに図1(b)ではより広い範囲の情報がなければ人の目にもいずれが正しいかさだかではない。言い換えれば各候補文字におけるあいまい性の程度とはその候補文字を通るパス(複数の場合もある)と、当該カラムにおいて他の候補文字を通るパスとの生起確率の比で表現するのが妥当であると考えられる。そこで各文字ごとのあいまい性の評価が可能な後処理として、次のような手法を採用した(図2)。

1.候補文字($S = Q_1,\ldots,Q_m$)について上記の拡張コスト付き形態素解析を行い、最適なパスのコストg_{opt}から一定値α以内のパスを求める。探索手法としてはDijkstraのアルゴリズムとビームサーチを併用している。

文 書 に つ い て　　発 明 の 詳 細 な 説 明
　 害　　　　　　　　　　　　　　　を
(a)　　　　　　(b)

図 1: 認識結果のあいまい性

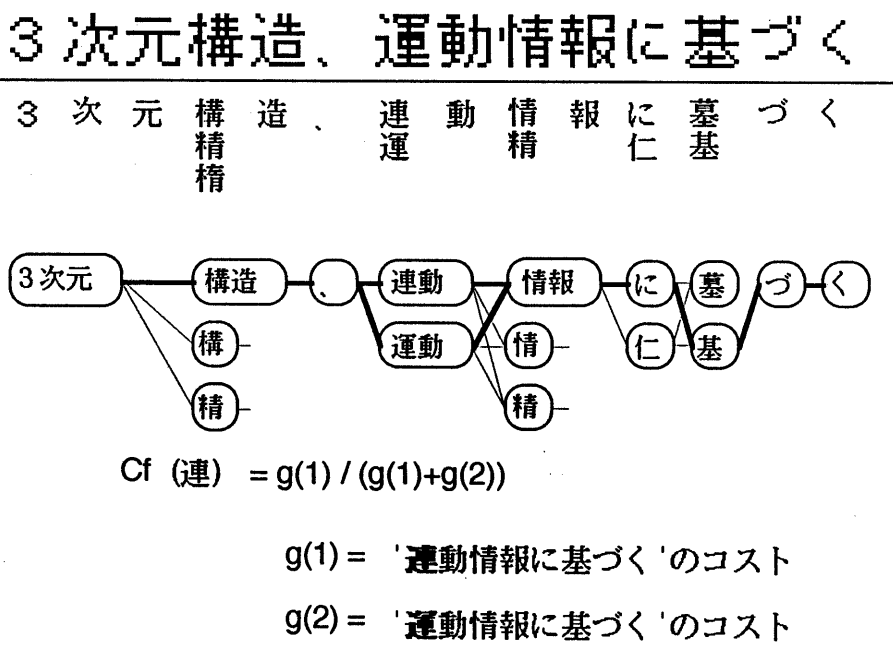

図 2: 複数パスの探索とあいまい性の評価

2.求まったパスを $P(i) = q_{i_1} q_{i_2}, \ldots, q_{i_m}$ $(i = 1, \ldots, N)$、さらに各文字位置 (j) について最適パスがその位置で採用した Q_j の要素を q_{opt}、最適パス同様その位置 (j) において q_{opt} を選択したパスを $P(i')$ $(1 \leq i' \leq N)$ とするとき、位置 j での確信度 (Cf) をつぎの式で定義する。

$$Cf = \sum_{i'} g(i') / \sum_{i=1}^{N} g(i)$$

ただし $g(i)(i = 1, \ldots, N)$ はパス $P(i)$ のコスト、左側の Σ は $P(i')$ に対応するコストの総和を求めることを意味する。

したがって Cf は (0,1] の変数で 1 に近いほど確信度が高いことになる。われわれのシステムでは各パスのコスト $g(i)$ は各単語の頻度、遷移確率、認識時の距離から得られる各文字の正解確率などの対数和で表現しているので Cf は文節単位の文脈を考慮したときの、当該文字の生起確率を近似していると考えることができる。たとえば図 2 の例では、可能なパスとして'三次元構造、運動情報に基づく'と'三次元構造、運動情報に基づく'の 2 つが残り、2 つのパスで文字が異なる 7 文字目で'連'と'運'に対する確信度が計算される。ここで複数パスが存在することと、あるカラムで複数の可能性があることは必ずしも一致しない。つまり'北大西洋'という複合語を例にとると('北大'という単語が辞書に存在する場合)'北大+西洋'および'北+大西洋'という 2 つのパスが出力されることになるが、文字列としては同一であり、認識結果という立場からみればあいまい性は存在しないからである。

3.Cfの値とq_{opt}が1位の認識結果と一致するか否かによりつぎのように候補の入れ換えや当該文字をMarking(候補入れ換え、警告)することによるオペレータへの通知を行う。

- q_{opt}が1位の認識結果と一致し、かつ$Cf > \delta$ならば何も行わない。
- q_{opt}が1位の認識結果と一致せず、かつ$Cf > \delta$ならば1位候補をq_{opt}に入れ替える。この場合もオペレータに対して(入れ替えたことを)通知する。
- q_{opt}が1位の認識結果と一致し、かつ$Cf \leq \delta$ならば警告を行う。
- q_{opt}が1位の認識結果と一致せず、かつ$Cf \leq \delta$ならば1位候補をq_{opt}に入れ替えた上で警告を行う。

3.3 キーワード抽出

前節で述べたように、本後処理はコスト付き形態素解析の拡張となっているので、後処理を行った時点で副産物として単語の切れ目、および品詞が分かることになる。そこで(複合語を含め)名詞を検出すればキーワードの候補が得られるがこれを表示する機能を付けている。これはデータベースへの文書入力を目的とする場合に必要な機能であるばかりでなく、最終的にオペレータが誤りを見付けるための補助手段としても重要である。

4 認識実験

本手法の効果を確かめるため認識実験を行った。用意したテストデータはA.電子情報通信学会論文誌(D分冊)の論文フロントページ(コピー: 計9455文字)、B.コンピュータに関する顧客研修用資料(ワープロ出力をオフセット印刷したもの: 計4129文字)、およびC.電気工学分野の特許公報(計2393文字)である。前2者は通常使用される程度の印字品質の代表として、Cは低印字品質の代表として比較的つぶれ、かすれが多く見られるものを選んだ。これらの文書にはコンピュータ関連用語が頻出するが、その多くはわれわれの自立語辞書に含まれていないため、ユーザ辞書に約300語登録した。後処理前の認識率と処理後の認識率および誤認識に対する検出率との関係をページ単位で図3に示す。ただし検出(図中Detect)とは当該文字に対して誤りの可能性があると識別し、候補の入れ替え、警告のいずれかが行なわれた(言い換えればMarkingされた)ことの意味で用いている。さらに対象文書ごとの平均値をとったものが表1である。本手法の効率を評価するため後処理後認識率の他以下に示す2つの尺度を用いる。

未検出率:検出されなかった誤認識文字の数(図3のUndetectedに相当) / 全文字数

総検出率:検出されたすべての文字数(過剰検出を含む) / 全文字数

表1: 対象別にみた後処理の効果

文書	認識率(%)		未検出率(%)	総検出率(%)
	後処理前	後処理後		
A	96.50	99.26	0.18	5.35
B	94.74	99.03	0.17	6.64
C	87.46	94.61	1.09	16.38

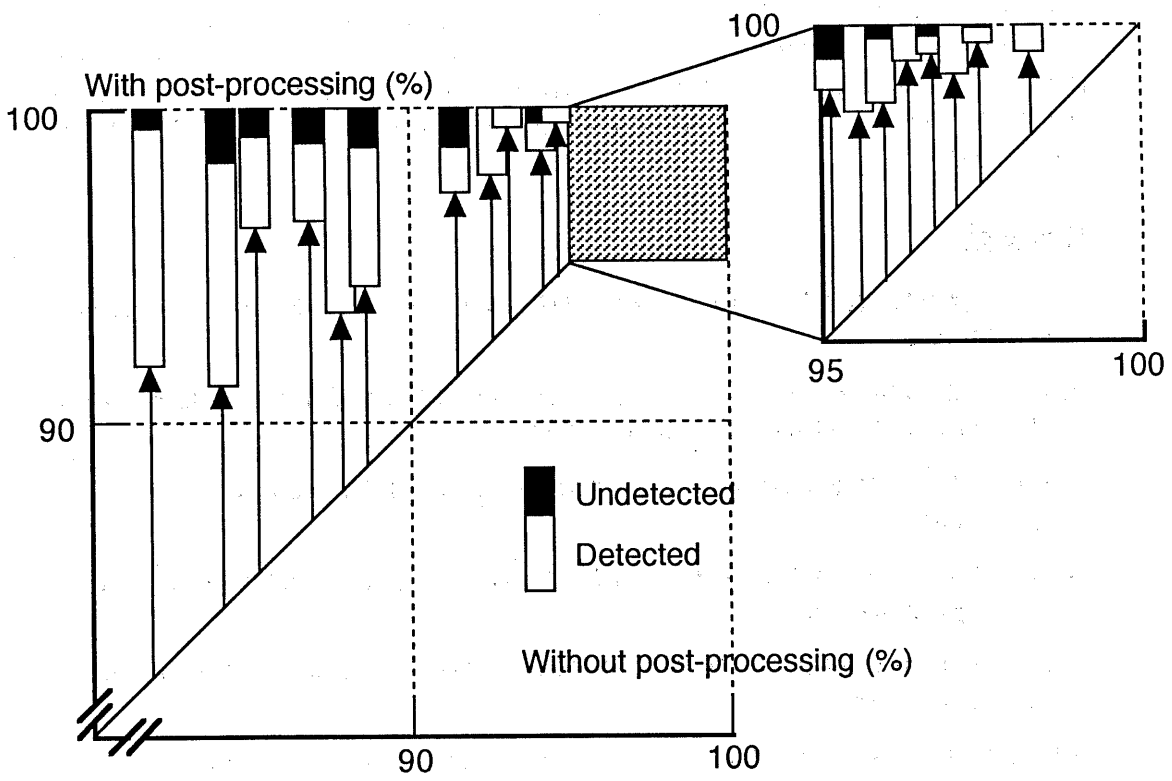

図 3: 後処理前の認識率と後処理後の認識率および誤認識文字の検出率

これらの図表から、後処理の効率は元の認識率に強く依存することが明らかであるが、後処理なしで 95%程度の認識率が確保できればそれを 99%程度まで引き上げ、かつ誤認識の見逃し (未検出率) を 0.2%程度に押えることができることがわかる。これはワープロ検定試験の 1 級が正解率 98.9%である [1] ことと比較すればほぼ十分な精度であると言える。さらにその場合総検出率が 5-6%にとどまっていることから、過剰検出も文節単位での処理としては十分少ないと考えられる。処理速度は第 2節で述べた環境で実行して約 27 文字/秒であった (ただし平均候補数は 2.3 個、最大候補数は 5 個の場合)。実際には本後処理が OS/2 上のマルチスレッドで実現されていること、および認識がカード上で実行されることから、オペレータは後処理のための待ち時間をほとんど意識することなく確認・修正が行える [6]。

5 まとめ

以上 DRS の文脈後処理機能とその実験結果について述べた。通常の使用環境においてはほぼ十分な精度、および速度で実行できる後処理機能が実現できた。特に後処理自身がその結果に対してあいまい性を評価し警告できることが (オペレータによる確認・修正を含めた) トータルな処理速度に貢献すると考えられる。ただし何らかの Marking(候補の入れ替え、または警告) が行なわれる率が約 5%(20 個に 1 個程度) というのはまだ多過ぎるという評価もできる。これについては過剰検出をさらに減らすと同時に、候補の入れ替えを行った場合でも十分な確信度ならば Marking しないことが良いと考えられるが未検出 (見逃し) との Trade off でありより多くの実験が必要と

考えている。また候補パスの選び方も現在は最良値から一定以内のコストをもつものという基準で選んでいるが、上位一定個数をとるということも考えられる。これについては探索手法と関連付けて検討が必要であろう[7]。今後は上記の課題他文書の対象分野を広げると同時に、後処理の有無によりオペレータの作業時間と最終的な入力精度がどのように変化するかを実験的に明らかにして行きたい。

参考文献

[1]宮原: 文書情報の蓄積検索システムに関する検討, 情処ヒューマンインタフェース研究会, 29-3, pp.1-10, 1990.

[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]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.

[4]池田, 大田, 上野: 手書き原稿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.

[5]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, Vol. 30, No. 11, pp.1394-1401, 1989.

[6]天野他: マルチメディア文書入力のための文書画像認識システム：ＤＲＳ, 情処マルチメディア通信と分散処理研究会, 48-6, pp.41-48, 1991.

[7]久光, 新田: 接続コスト最小法による形態素解析の提案と計算量の評価について, 電通言語理解とコミュニケーション研究会, NLC90-8, 1990.

[8]大河内: 仮名漢字変換のための形態素接続規則, IBM リサーチレポート N:G318-1560, 1981.