ヒューマンインタフェース 51－1
（1993．11．11）

文書処理システム用ドキュメントリーダ

伊藤悦雄 武田公人

(株)東芝 研究開発センター

大量文書データ入力支援用ドキュメントリーダを試作した。このドキュメントリーダは認識後の修正作業の軽減のためにHIと後処理を重視し、オペレータの修正支援機能や認識誤り自動修正機能などを有する。誤り自動修正では三文字間の接続情報を用いた。この接続情報は社内文書585文書（360万文字）を学習データとして抽出し、学習データとは異なる文書に出現する文字を他の文字に置換したデータで評価実験をした結果、10万文字の85.08%について元の文字列に自動修正することができた。また、文字種表示機能、認識誤りの前後の文字列を利用したかな漢字変換機能を有し、オペレータの修正作業を軽減することができた。

Document Reader for Documents Processing Systems

Etsuo Ito, Kimihito Takeda
R&D Center, TOSHIBA Corp.
Komukai-Toshiba-cho, Saiwai-ku,
Kawasaki, 210, Japan

We have developed a document reader for document processing systems, which has an automatic correction function based on extended trigram, as well as a kanji correction support function. The experiments showed that this automatic correction can recover 89.5% of the errors contained in 100,000 Japanese characters when 585 documents (3,600,000 Japanese characters) are used as training data. This document reader reduces the correction work required on the part of operators by displaying character types and eonverting kanji by referring to the characters preceding or folloing the errors.

1.はじめに

文書処理システムの発達に伴い、文字認識技術の用途が従来の帳票読取りから文書入力へ広がりつつある。文字認識では完全な認識は不可能であるため認識結果の修正の容易さが文書入力全体の効率を左右する[1]。このため、認識後の後処理を行うことによって読取り精度を向上させる方法として、認識された正解候補文字列に対して形態素解析を試みる方法[2-7]、文字間の接続情報で判別する方法[8,9]などが提案されている。

しかし、後処理による完全な誤り訂正は困難であるため、修正支援としてイメージと認識結果の対応表示、修正候補の一覧提示、修正の学習などが提案されている[10-12]。しかし、認識結果の修正に適した編集方式についての検討はあまりなされていない。

我々は、文書処理システムと接続した場合に文字認識システムに必要とされる機能ならびに後編集の効率向上のための手法を検討し、その結果に基づきドキュメントリーダを試作した。

本稿では、文書処理システム用ドキュメントリーダについて述べる。第2章で文書処理システムと接続した場合に文字認識システムに必要とされる機能の検討を行う。第3章では第2章で検討した結果に基づき試作したドキュメントリーダに取り入れた後処理およびHIについて述べる。

2.文書処理システム用ドキュメントリーダの要件

文字認識装置を用いた大量文書の入力では、認識結果の修正が入力全体の効率を左右する[1]。自動誤り修正能力を高めることが修正効率を上げる最も効果的な方法である。しかし、完全な自動修正は不可能なのでオペレータによる修正が必要である。

修正作業は誤りの識別と正答の入力に分類できる。誤り識別支援としてリジェクト文字や確度の低い文字の強調表示機能や検索機能を有する文字認識システムは多い。しかし、文字が誤読か正答かの判断や、誤読文字に対し提示される修正候補からの正答の選択はオペレータに任されている。そのため、オペレータに対し文字を識別するための支援機能はドキュメントリーダに必須である。

修正候補に正答が含まれない場合には正答をキーボードから入力するが、単語全体が誤っていることは少なく、1文字だけを修正する場合が多い。通常のエディタを用いてこれを修正する場合は、単語全体を入力し余分な文字を消去する必要がある。また入力の際は単語だけを独立して入力するので、かな漢字変換で通常行われる文法による変換候補の絞り込みを行うことができない。このため希望する文字を得るために余分な次候補要求を行う必要があり、修正効率が悪かった。従って、単語の一部だけを効率良く入力する正答入力支援が必要である。

次に、文書処理という観点から配慮しなければならない点について考察する。

文字認識システムでは基本的にデータを文字単位で扱うが文書処理システムでは文やパラグラフ単位で扱うことが多い。そのため、ドキュメントリーダにおいて文やパラグラフを正しく認識できる機能が重要である。また、機械翻訳やDTPなどの文書処理システムを考えた場合、文字列以外の情報例えばレイアウト情報を保存した形で文書を読み込める事が望ましい。したがって、レイアウト情報を抽出し、出力できる機能を有するべきである。

このように誤認識文字の自動修正機能、オペレータへの修正支援機能、文書処理システム用データ作成支援機能が文書処理システム用ドキュメントリーダには必要である。

本稿では以上のような考察に基づきドキュメントリーダを作成した。

3. ドキュメントリーダの試作

3.1.文字接続情報を用いた後処理

3.1.1 後処理方式

認識後処理として形態素解析を用いる方法[2-7]、文字接続情報を用いる方法[8,9]などが提案されている。

形態素解析を用いる方法では文字毎に複数の正解候補がある場合は膨大な組み合わせが発生する。全ての組み合わせに対して形態素解析をすることは困難であるため、単語の詳細分類やコスト付きの形態素解析、探索木生成方式による修正、品詞接続コストを用いた提案がされている[3-6]。また、全文字の全候補ではなく確度の低い文字に対して第二候補と入れ替えた文字列に対して形態素解析を行なうことにより高速化を図る方法が提案されている[7]。これらの後処理は専用ハードウエアを用いることによって十分な効果を発揮している。

一方、文字間の接続情報を用いる方法では、2文字単位の接続情報を用いるBigram、3文字の接続情報を用いるTrigramなどが提案されている[8,9]。この方法は形態素解析を用いる方法より高速に行うことができるが、Bigramでは「漢字-漢字」接続に対しては有効であるが「ひらがな-ひらがな」接続に対してはほとんどのかな同士が接続するためにあまり効果がないと報告されている[8]。そこで、筆者らは3文字単位の接続情報を用いるTrigramを採用する。

しかし、従来のTrigramに基づく方法では単語間の接続情報を用いていなかった。このため、一文字単語の誤りなど単語内では判断できない認識誤りや活用変化を含む認識誤りに対応できないという問題があった。そこで、ここでは単語間接続情報も用いることとした。接続情報辞書は実際の文書データに出現するすべての三連続文字から隣接出現度数を抽出して作成した。

3.1.2 実験・評価

上記方針に基づいた後処理の評価実験を行った(図1)。実験は社内文書データベース

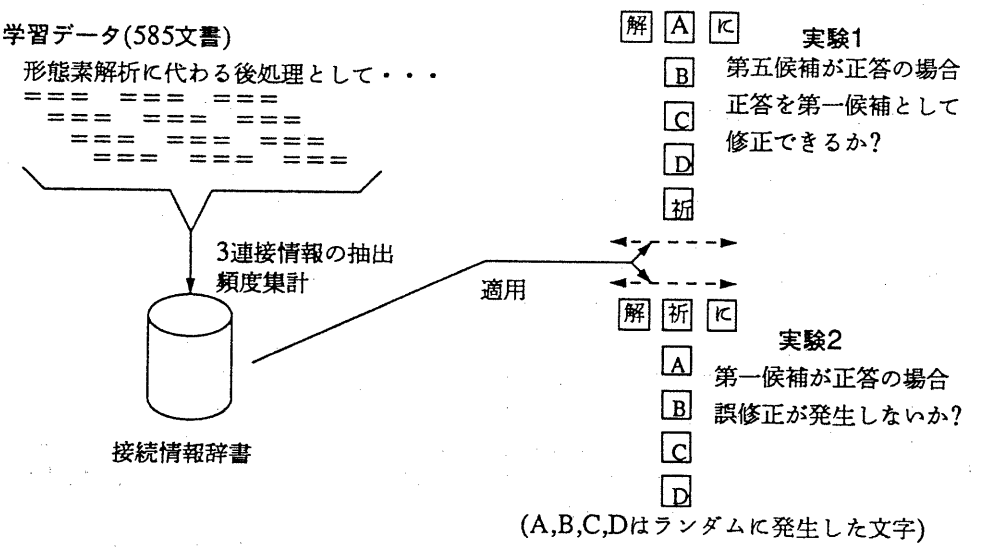

図1.後処理評価実験

を用い、学習データとして技術系社内報585文書(約360万文字)より抽出し、9万組の接続情報を得た。この接続情報辞書を用い、同じデータベースより学習データとは別の27文書(約17万文字)を用いて評価を行った。

誤った文字列をどの程度修正できるかを評価するために実験1を行った。実験1で用いたデータは三連続文字の中央の文字を他の文字に置換し、正答を第五候補としたものである。この結果、10万文字に対し85.08%が正しく修正できた(表1-a)。

次に、ランダムに発生した文字候補を与えることにより正しい文字列を他の文字候補に誤って置換してしまう割合を求めるために実験2を行った。その結果、10万文字に対し1.62%が誤って修正された。(表1-b)。

次に、この接続情報を用いて実際の認識結果に適用して測定を行った。文字認識には、当社で開発した複合類似度法を用いた[13]。認識に用いたデータは、学習データとは異なる技術系社内文書であり、認識率別の効果を検討するため、複写により文字の印刷状態を劣化させた原稿も使用した。この結果、54%の誤り回復が行えた(表2)。

実験1の結果と比べ回復率が半分程度に止まっている原因は、誤認識文字も前後の文字と接続可能である場合が多いこと、処理で用いる第五文字候補までに正答が含まれていない場合があること、連続する2文字が共に誤っている場合があること、英単語や数値など文書中に高い頻度で出現している英数字や記号が接続情報には含まれていないことなどである。

後処理により認識率が悪くなった例について検討すると、以下の2種に分けられる。

1) 正答と誤り文字列の両方が学習データ中に存在し、誤り文字列の方が出現頻度が高い場合
例:会計(合計)

2) 学習データ中に正答が含まれず誤り文字列の接続情報のみが含まれている場合
例:片カナ(片方ナ)

これらは、修正用の接続情報辞書における語彙数や接続情報が十分でない、あるいは原稿と学習データの文字使用傾向が異なることに起因する。したがって、大量文書から辞書を作成する、あるいは読み取る分野の文書による学習を行うことにより解決できる。

表1. 後処理実験結果

a) 実験1

回復	86,262文字(85.08%)
変化なし	15,126文字(14.92%)

b) 実験2

悪化	1,641文字(1.62%)
変化なし	99,747文字(98.38%)

全文字数 :101,388文字

表2.認識データに対する後処理結果

データ	認識率	誤り文字数	改善文字	悪化文字数
1	91.9%	51	31文字	3文字
2	95.4	18	10	2
3	98.4	4	1	0
4	98.7	14	5	1
5	99.2	9	5	0
6	99.3	6	3	0
計		102	55(54%)	6(6%)

・実験データ : 社内技術報告(計3690文字)
・認識率別の効果評価のため印刷状態を劣化させた原稿を使用

3.2 オペレータへの修正支援

3.2.1 文字判別支援

日本語には「鳥(とり)」と「烏(からす)」など類似した文字が多い。これらを区別するために、認識結果・候補を拡大表示し、オペレータによる識別の支援とする。しかし、例えばひらがなの「へ」とカタカナの「へ」など、拡大表示しても判別できない文字も存在する。認識した結果を人間が読むだけであれば、判別不能である文字が誤っていても文書としては通用する。しかし、機械翻訳など認識結果を電子的に処理する場合は、これらの誤りは許されない。そこで、判別不可能文字の区別のサポートとして文字種を提示することとした(図2)。

提示する文字種分類は、「ひらがな(清音)」「ひらがな(濁音)」「数字」など15種とした。これにより、同一文字種内には判別不可能文字が存在しなくなり、オペレータによる修正が容易になる。

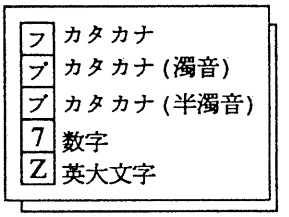

図2.文字種表示の例

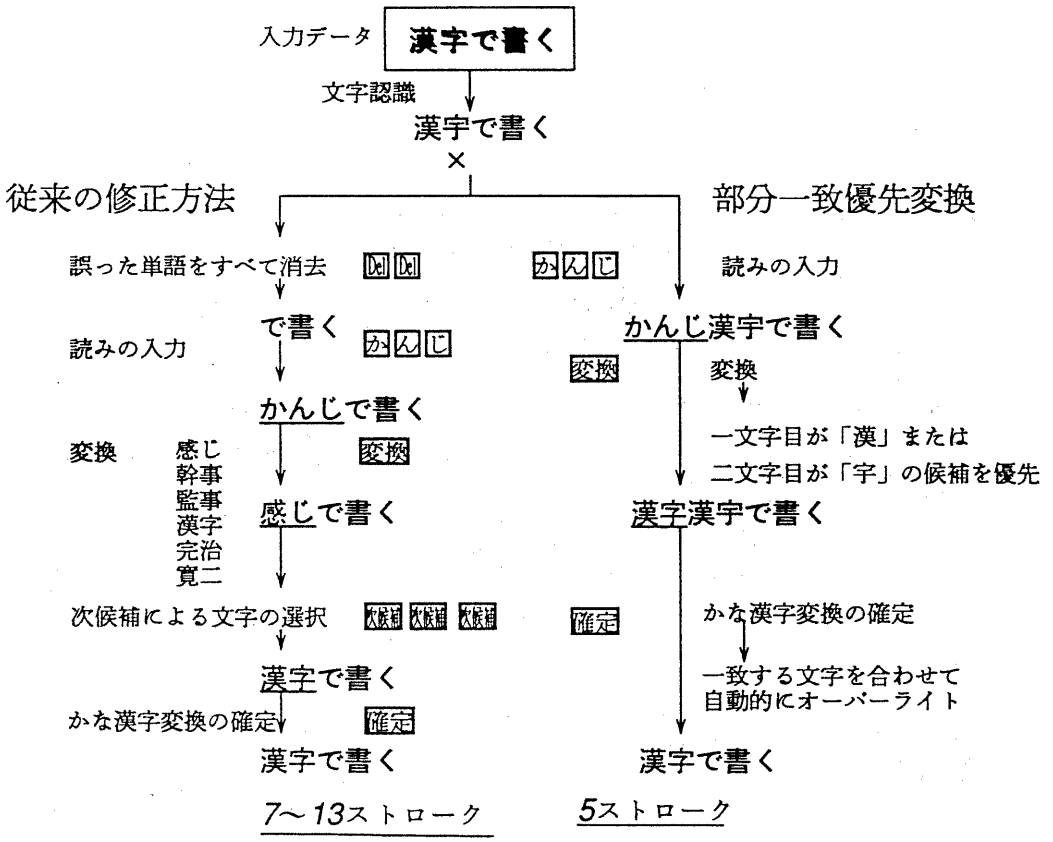

図3. 部分一致優先かな漢字変換

3.2.2 部分一致優先変換

誤認識文字の修正候補に正答が含まれない場合の正答入力支援として「部分一致優先変換」を考案した。

この入力方式は、認識において単語の一部のみが誤っている事を前提としている。キー入力により認識結果の修正を行う際、かな漢字変換で得られる同音語候補の内、修正する部分にすでに表示されている文字を含むものを優先し、変換の確定後は表示文字列と変換候補とで一致する文字を合わせて上書きするという方式である。図3に部分一致優先変換にゆる修正の例を示す。

例えば、「漢字」と認識されるべき文字が「漢字」と誤認識されたとする。この時、オペレータは文字入力カーソルをここに移動し、「かんじ」をかな漢字変換する。「漢字」「感じ」「幹事」「監事」などが候補として得られるが、本方式では「漢」の字が一致する「漢字」を第一候補として表示する。これにより無駄な同音語の選択操作をなくすことができる。ユーザが確定した際には「漢字」に「漢字」を上書きするため、誤り文字列を消去する必要がない。このようにかな漢字変換の次候補要求や文字列消去に必要な操作を削減でき効率良い修正ができる。

表2に示す認識結果において、全誤り102文字中、漢字2文字以上の単語中の認識誤りは65文字であった。その内、44文字は単語の一部のみの誤りであり、当方式の適用によって少ないキー操作で修正ができる。

3.3 文書処理システム用データの作成支援

3.3.1 「文」の認識率向上

従来の文字認識システムでは、認識結果は基本的に文字単位で扱われていた。後処理において形態素解析などを行う場合には文字列処理を行うが、この場合にも認識誤りによる解析失敗の範囲を限定するために、文ではなく句読点などで区切られた句を単位としてい

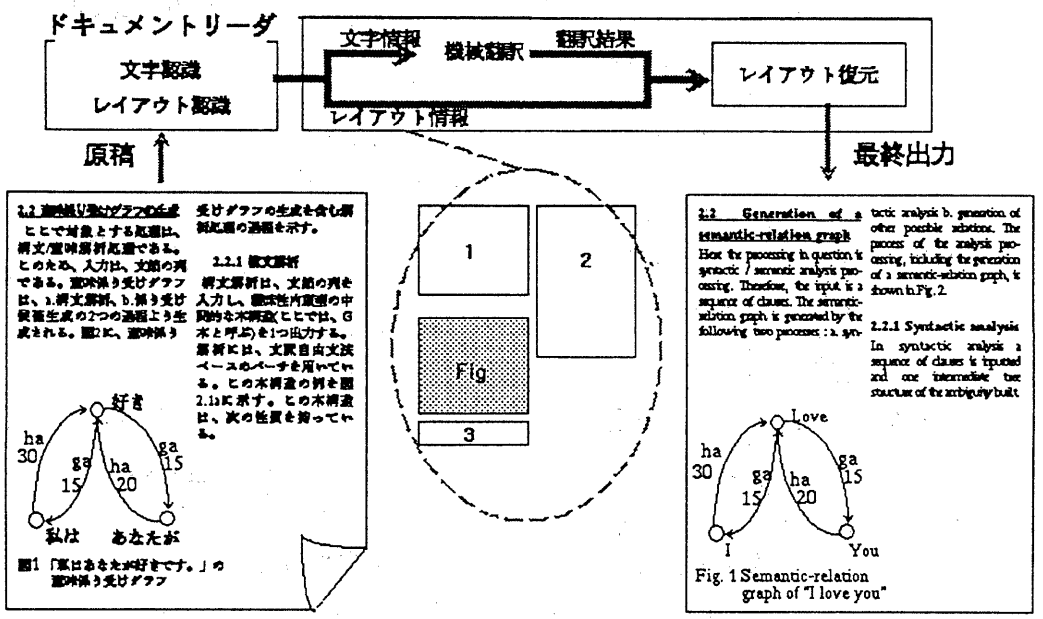

図4.機械翻訳システムを介したレイアウト情報の反映

る[6]。このため句読点の抽出も検討されているが、この場合にも句点と読点の区別には注意が図られていない[6,14]。

しかし、機械翻訳システムや要約システムなど文を単位として処理を行うシステムで認識結果を使用する場合には、文のセパレータある句点には特に注意を払う必要がある。また、句点以外にもタイトルや箇条書きなどでは改行でも文が区切れる点を考慮し、不要な改行の削除も必要である。

不要な改行の削除はパラグラフの終り以外の改行を削除することにより行う。これを文字情報だけで行うよりも、レイアウト情報やフォント情報などを利用した方が有効な処理となる。レイアウト情報を用いたパラグラフの境界の判定は、パラグラフ先頭行の開始位置は他の行より右(横書きの場合)か下(縦書きの場合)に寄っている、パラグラフの最終行は他の行より短いなどという基準で行える。

句点の認識率の向上のために、句点の前後の語句の解析を行う。句点の直前が終助詞か用言や助動詞の終止形である場合、あるいは直後が改行の場合には句点を優先する。

3.3.2 レイアウト情報の記録

機械翻訳システムでは、翻訳結果に原文イメージを反映させたいという要求が強い(図4)。

原文イメージはレイアウト認識機能を使用することにより、読み取ることができる[15]。本ドキュメントリーダでは認識したレイアウトをDTP形式に変換して出力する。

日本語の縦書きの原稿のレイアウトは、例えば読取り後、翻訳した場合、翻訳結果にはレイアウトが復元できないという問題点があった。しかし、これはイメージ回転という方法で解決した。

例えば、図5(a)に示す原稿を認識した場合、このレイアウトなどイメージ情報をそのまま保持し訳文である英文に反映させてしまうと、(b)のようになり反映させる意味が無い。そこで、レイアウト情報を回転させてから反映させることにより(c)となる。このレイアウト回転は、情報処理分野の論文、ビジネス・技術分野の雑誌のレイアウトにより正当性を検証した。

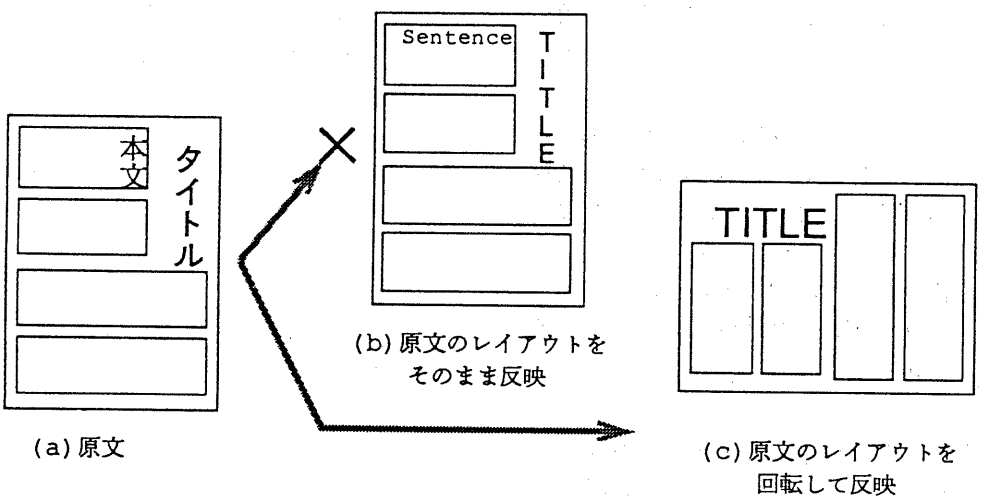

図5. レイアウト情報の回転反映

5.おわりに

文書処理システム入力用日本語ドキュメントリーダにおける認識結果修正時に有効な支援機能について検討した。その検討に基づき、認識誤りを自動修正できる後編集機能、漢字入力支援機能、文字種表示機能などを有するドキュメントリーダを試作した。

今後、試作したドキュメントリーダのHIの評価、原稿に含まれる図表やイメージ情報を取り込む機能を開発する予定である。

参考文献

[1] 宮原他、「文書情報の蓄積検索システムに関する検討」情処学会研究報告HI 29-3、1990

[2] 新谷他、「言語情報と認識情報を用いた文字認識後処理」信学技報PRL82-76、1983

[3] 久光他、「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」信学技報NLC90-8、1990

[4] 黒沢、「日本語文章を対象とする文字認識後処理方式」情処学会36回全国大会論文集、pp1801-1802、1988

[5] 伊東他、「OCR入力された日本語文誤り検出と自動訂正」情報処理学会論文誌、Vol.33,No.5、pp664-670、1992

[6] 杉村、「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」電子情報通信学会論文誌、Vol.j-72-D、No.7 pp993-1000、1985

[7] 山中他、「日本語テキストリーダにおける日本語構成支援機能」情処学会44回全国大会論文集、pp3-243 - 3-244、1992

[8] 杉村他、「文字連接情報を用いた読取り不能文字の判定処理」、電子通信学会論文誌、Vol.j-68-D、No.1 pp64-71、1985

[9] 小林他、「文字連接情報を利用した手書き文字列認識」情処学会研究報告NLC91-32、1991

[10]豊本他、「OCRエントリーシステムにおける訂正処理の生産性」情処学会42回全国大会論文集、pp2-154 - 2-155、1990

[11]宮原他、「日本語OCRにおける認識結果の一括修正」信学技報PRU93-84、1993

[12]村木他、「OCRの認識誤り訂正に於けるテキスト適合性の評価」信学技報NLC92-27、1992

[13]有吉他、「変形パターンの自動生成によるマルチフォント印刷漢字認識」、電子通信学会全国大会論文集、p.1465, 1987

[14]丹羽他、「文字認識後処理法と後処理による効果の分析」信学技報PRU91-135、1991

[15]S.Tujimoto, et al.. "Understanding Multi-Articled Documents" Proc. 10th Int. Conf. Pattern Recognition, Atrantic City, New Jersey, pp551-455, 1990