日本語OCR後処理のための確率主導型誤り検出法と文法主導型誤り検出法の比較

渥美 清隆　　増山 繁
atsumi@smlab.tutkie.tut.ac.jp　masuyama@tutkie.tut.ac.jp
豊橋技術科学大学 知識情報工学系

1 はじめに

近年, パーソナルコンピュータ上で動作する, 性能のよい日本語OCRシステム(以下, 単にOCRと省略する)が利用可能になりつつあるが, それに伴ない, 文書入力作業は軽減されたものの, OCR出力後の誤り文字の検出および訂正という作業にかなりの時間を割かなければならなくなった. ところが, このような誤り文字を検出する作業は長時間にわたる集中力を要求し, 人間に強度の負荷を与える作業であり, コンピュータによる誤り検出および訂正の支援が求められている.

このコンピュータによる誤り検出については, すでにいくつかの論文が報告されており [1, 2, 3, 4, 7], 活発な議論がされているが, それぞれの論文で提案されている検出法を客観的に比較評価した論文は少ない. 本論文では, 現在提案されている代表的な2つの誤り検出法として, 確率主導型の誤り検出法と文法主導型の誤り検出法について, できるだけ客観的な比較を試み, その性質について比較検討する.

2 非文の定義とOCRの誤りの性質

本研究における非文とは, OCRにかける前の原文と比較して, OCRから出力された文が1文字でも置換, 欠落, 誤挿入されているものを言う. 誤りの種類とその変換過程を表1にまとめた. 表1で示した変換の逆変換を誤り訂正演算であると定義するとき, いかなる非文も誤り訂正演算の組み合わせにより原文に復元することができる [5]

ところで, 今回使用したOCRでは, 文字の欠落誤りが単独で起ることはなく, 必ず置換誤りと複合して起きた. これは, ビットマップのイメージ上に何らかの情報が存在する場合には, それを必ず何らかの文字に変換するためである. ところが, 文字の大きさには依存せずに変換を行うため, 2文字以上の列を1文字と認識して変換してしまう場合がある. これは広義の意味での置換ではあるが, 本研究では1文字単位の誤り解析を行うため, このような誤りは欠落と置換が同時に起った複合誤りとする.

表 1: 誤りの種類

正しい文	誤った文	種類
$\cdots w_i\ w_{i+1}\cdots$	$\cdots w_i\ x\ w_{i+1}\cdots$	挿入
$\cdots w_{i-1}\ w_i\ w_{i+1}\cdots$	$\cdots w_{i-1}\ w_{i+1}\cdots$	欠落
$\cdots w_{i-1}\ w_i\ w_{i+1}\cdots$	$\cdots w_{i-1}\ x\ w_{i+1}\cdots$	置換

入力文を $w_0\ w_1\ w_2\cdots w_{n-1}\ w_n$,
$w_0,\ w_1,\cdots,w_n,\ x \in \Sigma$,
Σ は終端記号集合とする.

3 各誤り検出法の概説

我々の知る限りにおいて, 現在提案されている誤り検出法としては, 大きく分けて2つある. 1つはコーパスから得た統計情報を利用し, 誤りを検出する手法(以下, 確率主導型とする)[1, 2] と, もう1つは対象の言語を規定する文法を用いて, トップダウン的な解釈に基づいて誤り検出をする手法(以下, 文法主導型とする) [3, 4, 7] である.

しかし, これら2つのモデルとも, OCRの内部の情報を利用した誤り検出, つまり, 認識文字の第1候補のみならず, 第n候補までの認識文字を利用した誤り検出の研究報告が多い. 我々の立場としては, OCR以外のより広い誤り検出にも利用できるよう

にするため, 認識文字を第1候補文字のみ利用する, つまり, OCRがテキストを完全に出力した後に誤り検出をする方法を採っている.

ここでは, この第1候補文字のみを利用した, それぞれのモデルについての誤り検出法の概略について, 以下に述べる.

3.1 確率主導型誤り検出法

確率主導型の誤り検出法は更にいくつかの種類に分類できるが [1, 2], ここでは2重マルコフモデルを用いた誤り検出法 [2] について述べる. 2重マルコフモデルは大きく2つの部分に分けることができる. 1つは統計情報を得る部分であり, もう1つは統計情報を適用する部分である.

統計情報を得る部分では, 学習対象文字列 $x = w_1 w_2 \cdots w_n$ があるとき, 条件付き確率 $P(w_{i+2}|w_i w_{i+1})$ を計算する. 先頭文字と末尾文字に関してはそれぞれ特殊文字を用意して計算する. 例えば先頭文字では $P(w_1|\$\$)$, \$ は先頭を表す特殊文字, として計算する.

次に, このようにして得られた統計情報に基づき, 入力文字列 $y = v_1 v_2 \cdots v_n$ を解析する. 2重マルコフモデルでは, この入力文字列から任意の3文字列 $v_i v_{i+1} v_{i+2}$ を取り出し, 先に得た統計情報から条件付き確率 $P(v_{i+2}|v_i v_{i+1})$ を求める. このとき部分文字列 $v_{j-2} v_{j-1} \cdots v_{j+k} v_{j+k+1}$ について, 足切値 T を定めるとき, 連続して確率 P が T 以下であるならば, 部分文字列 $v_j \cdots v_{j+k-1}$ は誤り文字列であると推定する.

この推定方法は, 文献 [2] で述べられている単独の欠落誤りに対応することができないが, 2節でも述べたように, 今回使用したOCRでは単独の欠落誤りは出現しないので, この推定方法でも十分である.

本研究で作成した2重マルコフモデルで使用するための条件付き確率の統計情報は, 学習対象文字列 x を $\$\$w_1, \cdots, w_{i-1} w_i w_{i+1}, w_i w_{i+1} w_{i+2}, \cdots, w_n \&\&$ (\$ は先頭を示す文字を, & は末尾を示す文字である), のように, 3文字列の要素に分割し, 最初の2文字が同一であるような要素をそれぞれ1つに集めて, 条件付き確率の計算を行うことによって得た. 学習に際して, 学習対象文字列は段落単位で与えた. これは, 入力文字列の先頭文字や, 末尾文字が誤り文字であった場合に誤り文字を検出することが難しいため, やや大きなブロック単位で学習させることにより, 文の先頭文字や末尾文字の誤り検出もできるだけ行えるようにした.

学習用コーパスとしては, 日経新聞社が販売している1990年,1992年の新聞記事をまとめたCD-ROMから, 誤り検出用に用いる文を除いた社説のすべてと, 同じく春秋 (コラム) のすべてをそれぞれ用意し, 学習させた. また, この2つのコーパスを合せたコーパスからも学習させた. 社説は2年分で3.14MByte, 春秋は同じく2年分で0.93MByteである. これらから得た条件付き確率の項目数は社説で394409件, 春秋で210283件, 両方を合わせた場合では521946件を得ることができた.

3.2 文法主導型誤り検出法

文法主導型の誤り検出モデルもいくつかの種類に分類できるが, ここでは, 我々が研究を進めている誤り検出手法 [5, 6, 7] について述べる.

入力文字列 x を受理する言語 L(G) の文法 G が, 文脈自由文法 $G = (N, \Sigma, S, P)$ で表現できるとき, この文法を誤り訂正用文法 G' に拡張することができる [5]. 具体的な拡張方法は紙面の都合上述べないが, 2節で述べたような誤り変換の種類にそれぞれ対応した誤り訂正演算と呼ばれる導出規則を定義する. この導出規則を最小回数 [5], あるいは準最小回数 [6] 使用することで解析木を作成し, 入力列にいかなる誤りが含まれていても, 文法的には正しい解析木を得ることができる [5]. この方法を応用し, 誤り訂正演算が適用された部分文字列を, 誤りが含まれている部分文字列であると推定することによって, 誤りを検出することができる.

ところが, この方法は複数の解析木を出力するために, 誤り推定文字列を一意に定めることができない. そこで, 今回はこの複数の解析木のうち, 文節の区切りとして, もっとも適切な区切りになっている解析木を人間の手によって選択し, その解析木に含まれている誤り訂正演算を適用してる部分文字列を

誤りが含まれている部分文字列であると推定した.

本研究で作成した文法主導型誤り検出プログラムは主として, 文献[6]を採用し, 形態素解析としても動作するように拡張した[7]. 形態素解析として使用する辞書はICOTが提供しているTRIE辞書ユーティリティ[8]に含まれている形態素解析辞書(約15万件)を利用した. 文法は独自に作成し, 誤り訂正用文法に拡張した. 作成した文法の導出規則数は686個, 終端記号の数は366個であり, これを拡張した文法については, 導出規則数は2145個, 終端記号の数は366個で拡張する前と同じである.

4 実験方法

本論文における実験の目的は, 確率主導型誤り検出法と文法主導型誤り検出法の比較, 特に, 2重マルコフモデルの誤り検出法と誤り訂正演算を含む文法を利用した誤り検出法について, 比較をする. しかし, ここで行うのは性能の比較評価ではなく, 性質の比較評価であることに注意していただきたい. これは, マルコフモデルは, ある一定以上の大きさのコーパスを用意することにより, ある程度の性能を得ることができるが, 文法を用いた方法では, 辞書のチューニングや導出規則のチューニングに大きく左右されるため, 単純には性能比較ができないためである.

主に比較する内容としては, 社説と春秋のそれぞれ1遍ずつのOCR出力後の文書を用意し, その文書の誤り検出をそれぞれの検出法で実行する. 2重マルコフモデルの学習内容と足切値に関して, 社説については, 社説のみの学習を利用し, 足切値を変化させて性能を調査した. また, 春秋については足切値を0.001に固定し, 学習内容を春秋のみ, 社説のみ, 春秋と社説のそれぞれを用意し実験した. その結果について, 2種類の文書間でどのように誤り検出の仕方が変化するのかを見る. OCRにかける文書は, 600dpiのレーザープリンタから印刷し, それをイメージスキャナに読み取らせる方法をとった.

今回使用したOCRはBIRDS社製のThe OCR 日・英 Ver.1.0を用い, イメージスキャナはEPSONのGT-6500(300dpi)を利用した. このOCRに文書をかけた結果, 社説の方の記事は909文字を認識し, 内870文字を正しく認識した. 認識率は95.8%であった. 春秋の方の記事は677文字を認識し, 内622文字を正しく認識した. 認識率は93.1%であった.

表2: 社説について, 2重マルコフモデルの足切値による違いと文法主導型との比較

	2重マルコフモデル			文法型
足切値(T)	0.0005	0.001	0.002	
適合率[%]	52.5	52.4	50.7	18.1
再現率[%]	79.5	84.6	87.2	43.6

表3: 春秋について, 2重マルコフモデルの学習内容による違いと文法主導型との比較

	2重マルコフモデル			文法型
学習内容	春秋	社説	春秋と社説	
適合率[%]	23.2	21.2	29.6	19.0
再現率[%]	87.8	85.7	85.7	40.8

評価基準としては, 適合率と再現率を採用する. 適合率, 再現率はそれぞれ以下のような計算式に基づくものする.

$$\text{適合率} = \frac{\text{正しく誤りを推定した文字数}}{\text{誤りを推定したすべての文字数}} \times 100\ [\%]$$

$$\text{再現率} = \frac{\text{正しく誤りを推定した文字数}}{\text{実際に誤りであるすべての文字数}} \times 100\ [\%]$$

5 実験結果と両モデルの比較

実験結果を表2と表3に示す.

表2, 3の通り, 2重マルコフモデルの性能がかなり良いことが分る. 特に社説では, もとの記事の文体の統一性があるために, 実用レベルでの性能があると言える. ところが, 春秋の方では, 2重マルコフモデルについて, 再現率はある程度の水準を保ってはいるが, 適合率に関しては, 社説に比べて, いずれもかなり劣る. これは, 春秋というコラムの中での文体の統一性が社説に比べてかなり乏しいためと思われる. また, 学習内容が増加するに従って, 適合率は

上る傾向にあり, 再現率は下る傾向にある. これは, あまりに大きなコーパスから学習させた場合, 本来の目的である誤り検出が行えず, 誤り文字を見逃す可能性が増加することを示唆する. このため, どの程度の学習を行わせればよいのか, 十分に検討する必要があるだろう.

文法主導型の誤り検出は, 今回の実験では辞書や文法のチューニングが十分でなかったために, あまりよい結果は得られなかった. 主な原因としては, 辞書が形態素解析用であったために, 単漢字などが登録されており, これが再現率を大きく下げる要因となっている. しかし, 社説に対する結果と, 春秋に対する結果について, ほとんど性能差が見られなかったことに注意する必要がある. これは, 辞書が特定の分野に偏っていなければ, ある程度推敲した文書ならば, 文体に関わらず安定した性能を得ることができるのではないかと考えられる. 今回は, 2種類の文書でしか実験を行わなかったので, この性質については追試の必要がある.

最後に共通した性質として, 両検出法とも実際の誤り文字の直後を誤り推定文字とすることがかなりあった. これは, 何らかの誤りを発見してはいるものの, 前後の正しい文字に引っぱられて, 正しく推定することができなかったためである. このような誤り文字を検出するためには, 別の方策が必要となるかもしれない.

6 むすび

本論文では, 確率主導型の誤り検出法と文法主導型の誤り検出法の比較実験を行い, 以下のことを確認した.

- 2重マルコフモデルによる誤り検出は, 高い再現率を得ることができる.
- 2重マルコフモデルの学習内容や, 入力する文章の文体によって, 確率主導型の誤り検出は大きな性能差, 特に適合率の差がある.
- 文法主導型の誤り検出は文体によらず, 安定した性能を得ることができる可能性がある.
- どちらの検出法も誤り文字の直後を誤り推定文字とすることが多い.

今後は, 文法主導型誤り検出法の辞書や文法のチューニングを行い, 性能を高めるとともに, どのような文書にも安定して高い性能を得るために, 確率主導型誤り検出法と, 文法主導型誤り検出法の両方を組み合わせた誤り検出法についても検討予定である.

参考文献

[1] 森, 阿曽, 牧野 : "2重マルコフモデルを用いた日本語文書認識後処理", 情処研報 94-NL-102-12, Jul. (1994).

[2] 荒木, 池原, 塚原, 小松 : "マルコフモデルを用いたOCRからの誤り文字列の訂正効果", 情処研報 94-NL-102-13, Jul. (1994).

[3] 田邉, 木谷 : "文字認識誤り指摘のための形態素解析の適用性検討", 情処研報 94-NL-102-1, Jul. (1994).

[4] 黄瀬, 白石, 高松, 福永 : "構文・意味解析を用いた文字認識後処理法", 信学論, Vol.J77-D-II, pp.2199-2209, Nov. (1994).

[5] Aho and Peterson : "A Minimum Distance Error-Correcting Parser for Context-Free Languages", *SIAM J.Comput.*, pp.305-312, Dec. (1972).

[6] 渥美, 増山 : "構文解析上の自由度をもった非文訂正法の一提案", 信学論, Vol.J76-D-I, pp.686-688, Dec. (1993).

[7] 渥美, 増山 : "複数候補を出力する非文訂正法のOCR出力の誤り訂正への応用とその候補選出について", 情処研報 94-NL-104-4, Nov. (1994).

[8] (財) 新世代コンピュータ開発機構 : "TRIE辞書ユーティリティ" (1991).