# 文節文法を用いたイメージスキャナの読み取り結果の誤り検出

1Q－2

渥美清隆　増山 繁
豊橋技術科学大学

## 1　はじめに

近年, イメージスキャナを用いた文字認識システム(以下,OCRとする)がいくつか商品化されている. これらのOCRは, 文字を一文字毎に切り出し, 正規化してパターンマッチングを行なっている. 文字認識の精度は商品や環境によりまちまちであるが, だいたい95~98%を示している. この認識率はA4用紙1枚に1200文字程度書いてある文書を認識させた場合,1枚あたり24~60文字程度の誤りが生じることになる.

OCRが出力する文字の誤りも最終的に人間が訂正して正しい文にするが, 誤りの出現に規則性がなく, 訂正作業の負担は大きい. この作業量を減少させるため, コンピュータ支援による文の誤りを検出するシステムが望まれている.

最近の日本語の誤り検出及び訂正に関する研究としては, [1, 2] などが挙げられる. これらの研究はいずれも、単語あるいは文字の単位の前後の接続確率を考慮し, 尤もらしい検出または訂正を行っている. このような手法は, 誤りの特徴がはっきりしている場合や, 誤り訂正の作業が数種の単語あるいは文字の選択に置き換えることができるのであれば, 非常に有効な手段である.

しかしながら, 一般にOCRが出力する誤りには規則性がなく, またOCRから得られる情報を出力結果のテキストのみとしたため, 単語あるいは, 文字単位の接続確率を用いた手法は不向きであると考えられる.

そこで, 本研究では何らかの文法を導入することにより, 大局的な制限をうまく利用する誤り検出手法について考察する. 今回は, 文節文法を用いた形態素解析を行うことで, 誤り検出をする手法について考察する. この手法は, 形態素解析の手法として知られている最小単語法[1]に文節文法を組み込んだ手法で形態素解析を行い, 文節が成立するか否かを判定基準として誤り検出を行う. 本稿では, この手法に基づいたプログラムを作成し, 誤り検出に関する評価実験を行なったので, その結果について報告する.

An Error Detection on an Intra-Phrase Grammar for Japanese Text Read by Image Scanner
Kiyotaka ATSUMI and Shigeru MASUYAMA
Toyohashi Univ.of Tech.

[1]辞書に登録されている単語を最小個数使用する手法

## 2　文節文法の定義と句読点等の取り扱い

日本語文は文節のリストと句読点から構成されていることが知られている. 文節は次のような正規表現で定義することが出来る. これを文節文法と呼ぶ [3].

[文節] = [自立語][付属語]*
[自立語] = [接頭辞]*[語幹][活用語尾]*[接尾辞]*

この文節文法を実際に適用する場合には, もう少し複雑な構文規則を用意しなければならない. 例えば語幹と活用語尾との組み合わせには制限があり, それを制御するための構文規則を書かなければならない等である.

また, 日本語表記に出現する記号のうち, 開き括弧と句読点 (‘。’,‘、’など) は疑似文節として扱う. 閉じ括弧については, 疑似自立語とする. このようにする理由は, これらの文字をプログラム中で特別に扱うことなく, 文法の書き換えだけで対処するためである.

## 3　実験

本実験の目的は, 作成した誤り検出プログラムがどの程度の誤り検出を行うことができるのかを見る. 実験用文章としては, 日経サイエンス1993年7月号 pp.66~76「核-マントル境界領域」を使用した. これを対象として取り上げる理由は, 文字認識を行なう最適の紙質であり, 文字の大きさも適当だからである. この文章は11744文字であるが,OCRに入力した結果,11790文字を認識し, その内誤った文字数は452文字であった.OCRはBIRS社製のThe OCR 日・英を使用した.

この文章には高々452文字の誤りしか含まれていないが, その誤りの特徴は1節でも述べたように, ほとんど見受けられなかった. このような文章に対して, 如何に適切な誤り検出が行なえるのかを考察する.

### 3.1　実験方法

まず, 実験の評価基準として, 適合率と再現率を導入する. 適合率とは, 誤り検出プログラムが指摘した誤りの可能性がある部分に対して, どの程度正しく誤りを指摘したかであり, 再現率とは,OCRが出力する結果に含まれている誤りを, どの程度指摘したかであると定義する. これらを式に表すと次のようになる[2].

[2]文献 [2] などでは文節単位の評価であるが, 本実験では文字を単位として評価する.

$$適合率 = \frac{正しく誤りを指摘した部分の文字の数}{誤りを指摘した全ての文字の数}$$

$$再現率 = \frac{正しく誤りを指摘した部分の文字の数}{全体の誤りの文字の数}$$

実験は, 予めOCRによって得られた結果と, それを人間が正しく校正した文章を用意し,OCRの結果の文章を誤り検出プログラムにかけて, 校正した文章と検出プログラムが出力した内容を比較して, 評価を行なう.

検出プログラムのアルゴリズムとしては, 最小単語法に文節文法を組み込んだ手法を採用した. これは, 辞書にある単語以外でも文節文法によって導出可能であれば, 導出した結果も1単語とする方法である. その結果文節が構成出来ればその文字列は1単語とするが, これにより, 文節を構成できなかった部分は誤りを含む部分であるとした. 文節文法として用意した文法は, 終端記号62種類, 導出規則53個を用いた. プログラムはC言語で記述し, 形態素解析に必要な辞書は岩波の計算機可読の辞書 (15万語) を使用した.

## 3.2 実験結果

実験の結果を表1の実験結果に示す. この中の誤り的中文字数とは誤っているとした文字の中で, 実際に誤りであった文字の数である. 誤り検出文字数とは, 文節を構成できなかった部分のすべての文字の数である.

## 3.3 実験結果の改善

実験結果として, 適合率については,1文字の誤りを推定する場合でも, 平均文節長と同じ位の文字数を同時に誤りを含む文字列として出力するため, この程度でも止むを得ない. しかし, 再現率については, 実験用プログラムが出力する文節として構成出来ない部分, つまり誤りであると推定した部分を数え上げただけでは, 表1の通りあまり良い結果とは言えない. そこで, この結果を改善するため, 次のような仮定を考えてみる.

仮定1 1文字で文節を構成する事はない.

仮定2 本システムの処理で文節を構成できない部分は, 隣接した文節に誤りがある.

この仮定に基づき, 次のような誤りの推定の改善と, その評価を考えてみる. 仮定1からは, 実験結果のうち1文字で文節を構成しているものは誤りであると推定する. 仮定2からは, 実験結果で誤りであると推定した部分に隣接している, 文節として成立しているものを誤りであると推定する. そして, 仮定1と仮定2を組み合わせたものについて考える. それぞれの結果を表1の仮定1, 仮定2, 仮定1,2, にまとめた.

| | 実験結果 | 仮定1 | 仮定2 | 仮定1,2 |
|---|---|---|---|---|
| 誤り検出文字数 | 729 | 1106 | 2114 | 3489 |
| 誤り的中文字数 | 157 | 243 | 211 | 341 |
| 適合率 (%) | 21.5 | 22.0 | 10.0 | 9.77 |
| 再現率 (%) | 34.7 | 53.8 | 46.7 | 75.4 |

表 1: 実験結果

## 3.4 改善結果の考察

先の改善結果から, 仮定1及び仮定2と, その組み合せは, 再現率を向上させることが分った. 特に, 仮定1と仮定2の組み合せでは適合率が75.4%と大幅に向上した. しかしながら, 仮定2を併用すると, 誤りであると推定する文字の範囲が大幅に広がり, 適合率は下った.

# 4 今後の課題

現在の方法では適合率が低いので, 再現率を高い水準で維持したまま, 適合率を向上させることが当面の課題である. しかしながら, 既に文節として成立している部分に対して, 新たに誤りを含む部分であると推定するには, 後述するような文節構造以外の他の情報が必要になる. また, 再現率が最も高い方法でも,25% 程度の誤りを見逃がしていることになるが, この見逃がしている部分のほとんどは, それ単独の文字列として見れば文節と見えてしまうためである.

このように, 一見正しいように見える, 誤りを含んだ文字列に対しても誤りを推定出来るようにするため, 文節単位の構文解析による係り受けを調査し, 係り受け構造の矛盾を発見する手法と, シソーラス情報を用いて, 段落単位で中心になる意味分類から外れるような意味を持つ文節を, 誤りを含む文字列であると推定する方法について検討したい.

# 参考文献

[1] 下村 秀樹, 並木 美太郎, 中川 正樹, 高橋 延匡. 最小コストパス探索モデルの形態素解析に基づく日本語誤り検出の一方式, 情報処理学会論文誌, Vol.33, No.4, pp.457-464(1992).

[2] T.Araki,S.Ikehara and N.Tsukahara. New Methods for Deciding types of Erroneous Characters Wrongly Substituted,Deleted and Inserted in Japanese Bunsetsu and Correcting These Errors,*Proceedings of NLPRS '93*,pp.101-108(1993).

[3] 長尾 真監修. 日本語情報処理, 電子通信学会 (1987).