3K－8

文脈自由文法上のある非文訂正技術を用いたOCR出力後の誤り検出

渥美 清隆　atsumi@smlab.tutkie.tut.ac.jp
増山 繁　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

イメージスキャナなどを用いた文字認識技術は年々向上し，その需要も増えつつあるが，文字認識後の複数候補の選択や，誤って選択した文字を検出する機構が，現在のところ十分に発達しているとは言い難い．このような誤り検出をするための研究として，隣接文字間の接続確率を用いた方法[1]などが成果を上げているが，確率の学習を十分に行なわせるためには，かなりの量のコーパスが必要であり，実用段階には至っていない．また，最小文節法に基づいた方法[4]では，誤り推定の候補を1つだけしか出力しなかったため，誤り検出に十分な能力を示してはいない．

一方，文脈自由文法上の非文訂正法として，文献[2, 3]などが提案されている．これらの方法は，元の言語を文脈自由文法上の文法として規定できれば，それを誤り訂正用の文法に変換することにより，誤った文が入力されても，正しい文が入力されたときと同じように解析可能となる．しかもこれらの方法は，どこに誤りがあったのかも知ることができる．特に我々が開発した方法[3]（以下，渥美らの方法とする）は，後で示す誤り訂正演算の使用回数が最小の場合だけの解析に留めず，最小から最小$+d$までの誤り訂正演算の使用回数までの解析をすることにより，かなり広範な誤り訂正ができるものと期待できる．このような方法が現在まであまり注目されていなかったのは，入力列の長さをnとするとき，実行時間が$O(n^3)$もかかることや，使用する記憶容量が$O(n^2)$になり，採用が困難であったためだが，近年の計算機の高速化，大容量化に伴ない，十分実行可能な環境が整ったと考えている．そこで本稿では，渥美らの方法を形態素解析に応用した誤り検出手法を提案し，その手法について計算機実験を行ったので，それを報告する．

2 非文の定義

本研究における正しい文（以下，正文とする）とは，OCRシステムにかける前の原文であり，誤った文（以下，非文とする）とは，OCRシステムにかけた後の出力文章のうち，原文と1文字でも異なった文字がある文であると定義する．このような非文は表1のような3種類の誤りの組合せにより，正文から非文に変更した文であると見なすことができる[2]．

表1: 誤り訂正演算の種類

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．

この3種類の演算を誤り訂正演算とする．最小誤り訂正演算回数とはこれら3種類の誤り訂正演算を使用して，入力された非文から出発して最も近い正文にたどりつくのに必要な最小限の誤り訂正演算の使用回数のことである．また，許容度dの準最小誤り訂正とは，入力された非文からある正しい文に最小回数$+d$以内でたどりついた時をいう．

3 文節文法の定義と誤り訂正文法への変換

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[5]．

[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．

さらにこの文法から，誤り訂正文法へ変換[2]を行う．誤り訂正文法への変換手順については文献[2]に詳述されているので省略するが，表1に示した誤り訂正演算をうまく導出規則の中に取り込むようにしている．今回評価実験のために用意した文節文法の導出規則は54個，品詞に相当する終端記号は151個であり，誤り訂正文法へ変換した結果，導出規則は302個となった．

An Error Detection for Japanese Text Read by OCR Systems Based on an Error-Correcting Technique for Context-Free Grammar
Kiyotaka ATSUMI and Shigeru MASUYAMA
Dept. of Knowledge-based Inform. Eng., Toyohashi Univ. of Tech.

4 非文訂正法の形態素解析への拡張

文献 [2, 3] で述べられている非文訂正法は, Earley 法に基づいた文脈自由文法上で適用できる方法のため, 形態素解析に利用するためには次の問題点を解決する必要がある.

- 文字単位を終端記号として解析するには, 形態素用の辞書をすべて導出規則として展開しなければならないので, 記憶容量上の問題により困難.
- 形態素の品詞単位を終端記号として解析するには, 形態素の区切りが複数候補存在するために, 入力列の長さ (形態素数) すら, 定義することが困難.

そこで, Earley 法の終端記号からの導出の部分を拡張することにより, うまく形態素解析が行なえるようにした. 以下では, Earley 法の概略を説明し, 終端記号からの導出の部分の拡張について説明する.

Earley 法では解析テーブルという集合の枠を $n+1$ 個用意し, そこに解析木の節点に相当する要素を生成することにより解析を行う. Earley 法によって生成される要素は $[A \to \alpha \bullet \beta,\ p]$ という形をしている. このとき各記号の意味は, 終端記号の集合を Σ, 非終端記号の集合を N とするとき, $A \in N$, $\alpha, \beta \in \{N \cup \Sigma\}^*$, $A \to \alpha\beta$ は導出規則である. $\bullet$ はそこまでの部分解析木が完成していることを示すメタ記号, p はこの要素が何文字目から解析を始めたのかを示す時点ポインタである.

さて, 入力された文字列 $x = w_1 w_2 \cdots w_n$ があり, w_i 文字から w_j 文字までが品詞 a に置き換えられるとする. このとき終端記号からの導出について, 解析テーブル I_{i-1} に $[A \to \alpha \bullet a\beta,\ p]$ なる要素が存在するならば, 解析テーブル I_j に $[A \to \alpha a \bullet \beta,\ p]$ を生成する. 通常の Earley 法では, 終端記号は 1 文字ごとにしか解析できないが, ここで拡張した方法を使うことにより, 複数文字列が 1 終端記号として扱われる場合も解析可能となり, 形態素解析にも利用できるようになる.

5 実験

今回提案した誤り検出手法がどの程度の能力を持つのかを検証するために, 実際に計算機実験を行った. 実験方法は誤りの含まれる文を実際に入力し, 許容度 1 の準最小誤り訂正の範囲の誤り推定候補を出力させ, その候補数, および正しく誤りを推定した候補数を測定した. 実験用の入力文は日経サイエンス 1993 年 7 月号 pp.66~77「核 - マントル境界領域」を実際の市販のパソコン用 OCR システムに入力した結果から得られた, 誤りを含む 8 文を抜き出した. 8 文中には 151 文字が含まれ, そのうち 11 文字が誤りであり, すべて置換誤りであった. これらの文を今回作成したプログラムに入力した結果を表 2 に示す. 文 8 の B は, すべての誤字について正しく指摘した候補数であり, もし, 1 つでも正しく誤字を指摘した候補数とするならば, 90 となる.

表 2: 実験結果

文	字数	誤字数	候補数	A	B
1	21	1	31	3	31
2	23	1	45	0	11
3	16	1	120	0	22
4	15	1	138	1	122
5	16	1	41	0	19
6	17	1	45	0	27
7	22	1	121	0	40
8	21	4	90	1	24

候補数は誤りを推定した候補数, A は最小誤り訂正演算回数, B は誤りを正しく指摘した候補の数を示す.

6 考察

実験の結果, 最小誤り訂正演算回数が 1 以上であるときは, かなりの確率で誤字を正しく指摘した誤り推定候補を選択できるが, 最小誤り訂正演算回数が 0 のときでも正しく指摘した誤り推定候補が存在する. 文献 [2] の方法では, このような非文を検出することは困難であったので, 渥美らの方法が有効に動作することが分る. しかし, 最小誤り訂正演算回数 +1 までを許容したことにより, 誤字の指摘をすることが可能になったかわりに, それ以外の候補もかなり多く出力してしまった.

今後は, このような誤字を正しく指摘できなかった誤り推定候補を削減するために, 文節間の接続に関する文法 (拡張格文法など) を導入したい. また, 誤り検出を行った部分に対する, 訂正についても検討したい.

参考文献

[1] T.Araki, S.Ikehara, 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.

[2] A.V.Aho, T.G.Peterson: A Minimum Distance Error-Correcting Parser for Context-Free Languages, *SIAM J.Comput.*, pp.305-312, Dec., 1972.

[3] 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.

[4] 渥美, 増山: 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, 情処第 48 回全国大会, Vol.3, pp-39-40, Mar., 1994.

[5] 長尾: 日本語情報処理, 電子通信学会, 1987.