OCR 自動補正用分野別辞書の検討

2L－5

江澤 義典
関西大学

嶋岡 和章
関西大学 大学院

1 はじめに

光学式文字読み取り装置 OCR (Optical Character Reader) は, いわゆるパターン認識機械の一つであり, 印刷された日本語文書を電子化する道具として大変有力である. しかしながら, 現段階においては OCR の文字認識率を 100% にまで向上させることは不可能であると言われている. また, OCR の誤りは対象文書の内容とか形式に依存して著しい偏在を示す傾向が指摘されている [1]. そして, 文字認識の誤りを補正する後処理の研究が, 数多く行なわれている [2][3].

とくに, 伊東と丸山による DRS[3] は汎用の辞書 (115,900 語) に加えてユーザ辞書 (1000 語) を認識実験用に用いた本格的なものであるが, OCR 装置のアルゴリズム自身に改良を施そうとするものである. すなわち, 認識過程の曖昧性を表現する方法として単語の出現頻度, 単語間の遷移確率, 認識実験確率および候補生起確率を基にした確信度を導入しているので, この方法は元になる OCR の認識アルゴリズムに依存したものであり, 独立した補正法には採用しがたい.

本研究では, 既存の OCR 装置を利用して, その後のオペレータによる補正処理 (後処理) を自動化するための辞書を構築する方法について検討する. とくに, 対象分野を刑法テキストに限定した場合に, 約 3% の誤りを含む OCR 出力文書に対して補正率を向上させ, 過剰な変更を極力押さえる方法を検討した.

2 OCR 自動補正システム

本研究における OCR 文書の自動補正システムについてその概要を述べる.

Dictionary Customizing for Error-Correction of the Japanese OCR Outputs
Yoshinori Ezawa, Kazunori Shimaoka
Faculty of Engineering, Kansai University
Kansai University Graduate School
3-3-35 Yamate, Suita, Osaka, 564, Japan

2.1 自動補正方式

OCR 装置としては FMR の上の「漢字 IOCR」を用いた. そしてワークステーションにおいて以下の補正処理を行うシステムを構築した.

Step_0 OCR 文書の適当な部分列 (トークン) を抜き出す.

Step_1 過剰変更停止辞書に登録されているか否かを調べる. この辞書に登録されておれば Step_5 へ.

Step_2 補正文字列辞書の登録語であるか否かを調べる. この辞書に登録されていなければ Step_5 へ.

Step_3 補正文字列辞書の訂正語と置換する.

Step_4 置換した結果が望ましくない場合には過剰変更停止辞書または補正文字列辞書を更新する.

Step_5 文書の終わりまで Step_0 から繰り返す.

本方式は対話型の補正を実現しているが一括方式で補正処理する場合には上記の Step_4 も一括処理することになる.

2.2 辞書の構築

OCR の間違いは偏在する傾向があり特定の文字列が繰り返し同じように誤って認識されるという性質を利用して, テスト文書を数回処理し, その結果を用いて対話的に辞書を構築する. もちろん, 一括処理による辞書構築も可能である.

補正文字列辞書 補正すべき誤った仮名漢字文字列と対応する文字列とを登録しておく.

過剰変更停止辞書 補正すべきではない仮名漢字文字列を登録しておく. 既存の国語辞典とか漢字辞典を参照する.

2.3 辞書の更新

新規文書に対して新たに必要な語句を辞書に登録する.

3 辞書構築の実例

対象分野として司法試験問題集の刑法分野を取り上げた場合について報告する.

3.1 刑法テキストの特徴

刑法分野では, 通常の日本語文と比較して, 独特な漢字列が多く含まれている. (例：禁錮, 懲役, 罰金, 撲殺, 事件, 尊属, 湮滅など)

3.1.1 共通変更文字列辞書

どの分野にも共通して現われる漢字は常用漢字の範囲と考えられるが, そこでも OCR の認識誤りは発生する. 本実験では 101 個の漢字を登録した. (例： 因→国 , 下→不, 金→余 など)

さらに, どの分野にも共通して現われる仮名文字の OCR 認識誤りも発生する. 本実験では 99 個の仮名文字列を登録した. (例：なかて → なかで, しそも→しても, しがし、→ しかし、等)

3.1.2 共通過剰変更停止辞書

過剰変更を抑止するために過剰変更停止辞書を用意する必要がある. (例: 因果, 原因 など 25 個の単語とか落下, 地下, 下見, 下車 など 72 個の単語などがある [4].)

3.2 分野別辞書の更新

刑法分野に固有の辞書として登録した変更文字列はわずかに 18 字であった. (例：犯, 懲, 錮, 湮など)

3.3 自動補正実験

まず, 文字数 9495 字の司法試験問題集の一部（刑法理論, 刑罰法規の解釈など）を OCR 装置で処理し, そのときの誤りを対話的に検出して分野固有辞書とした. このとき, スキャンのタイミングによって誤差が発生するので, 全く同一の原稿を 3 度処理させ (1.1, 1.2, 1.3) それらの認識結果を総合して辞書を構築した. つぎに, 犯罪論および構成要件などの部分 (19571 字) を OCR で処理した結果 (2.0) を補正してみた.

No.	1.1	1.2	1.3	2.0
文字数	10322	10641	10682	21830
漢字誤り数	52	51	49	315
仮名誤り数	158	166	177	339
認識率	0.979	0.979	0.978	0.970
漢字補正率	0.981	0.961	0.979	0.508
仮名補正率	0.999	0.999	0.999	0.640
総合正解率	0.999	0.999	0.999	0.987

4 おわりに

対象文書の分野を限定したときに OCR 出力文書の自動補正が可能となる辞書の構築方法について検討した. そして, 刑法テキスト分野における実験結果より非常にコンパクトな分野別辞書を準備するだけで, 通常の使用環境では十分な精度の自動補正が可能であることが分かった. 認識率 97%の文書を補正して 98.7%の正解率を得た. 今後は, 民法とか民事訴訟法などの分野別辞書の検討を行いたい. また, 形態素文脈情報を利用した補正法 [5] についても検討したい.

謝辞　日頃からご討論いただく, 伊藤教授および植村教授に深謝いたします. なお, 本研究の一部は関西大学法学研究所の法学教育研究班の共同研究として実施されたものである.

参考文献

[1] 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. J72-D-II, No. 7 (1989).

[2] 新谷, 梅田：文字認識における複合後処理法の能力評価, 電子情報通信学会論文誌, Vol. J68-D, No. 5 (1985).

[3] 伊東, 丸山：OCR 入力された日本語文の誤り検出と自動訂正, 情報処理学会論文誌, Vol. 33, No. 5 (1992).

[4] 貝塚, 藤野, 小野：角川漢和中辞典, 角川書店 (1991).

[5] 下村, 並木, 中川, 高橋：最小コストパス探索モデルの形態素解析に基づく日本文誤り検出の一方式, 情報処理学会論文誌, Vol. 33, No. 4 (1992).