負データが極端に少ない訓練事例を用いるOCR誤認識検出

田中 大輔† 田村 直良††

†横浜国立大学 教育人間科学部 マルチメディア文化課程
††横浜国立大学大学院 環境情報研究院
{dtanaka,tam}@tamlab.eis.ynu.ac.jp

1 はじめに

本稿では、OCRシステムによって認識された文章中の誤りを機械学習により検出する手法について論じる。

近年、官公庁における情報公開や、新聞記事データベースに見られるように、膨大な文書をインターネットで閲覧できるようになってきている。これらを実現するために、紙媒体の文書を情報検索に利用できるテキストデータに電子化することが要求されている。一般に文書をテキストデータに電子化するには、イメージスキャナにより紙媒体の文書を画像化し、認識する文字領域を選択し、OCRによって文字認識を行う。

現在、市販されているOCRエンジンは、ノイズの少ない原稿では99%前後の認識精度を実現している。しかし、一般的には、誤り文字が含まれていることが気にならない認識精度は99.95%以上が必要とされている。そのため、誤認識を訂正する何らかの後処理が必要である。後処理は人手に依る作業が大部分を占め負担が大きい。そこで、これら人手による作業の補助となるシステムが望まれる。

今までの研究では、誤認識の訂正を行っているものが数多い。自動的に誤認識を訂正することは、正しい文字までをも改悪する弊害を持ち合わせている。そのため、99.95%以上の精度を満たすことができていない。竹内ら[6]は共起情報を用いて検出を行い、精度の高い文字の修正を行っている。共起情報は誤り文字の含まれない文書から文字tri-gramモデルとして作成している。そして、OCR出力した文字がその文字tri-gramモデルに含まれていなければ誤りであるとしている。しかし、誤り文字列は常に、日本語として存在しない文字列であるとは限らない。つまり、誤りのない文章のtri-gramモデルに合致するOCR誤認識文字を含む文字列が存在する。よって、そういった文字trigramモデルに含まれる誤り文字は検出される可能性が低いと考えられる。また、久光ら[3]はOCRの文字候補付きの出力結果と、形態素解析を用いて後処理の効率化を図っている。この手法ではOCR認識エンジンの精度に依るところが大きい。

これらの他にOCR誤認識文字検出及び訂正に関するさまざまな研究が為されているが、多くは95%以下のOCR認識精度を前提にしている。本研究では、99%程度の精度の認識結果に対して誤認識文字を発見して、99.95%以上の精度の文章を得ることを目標としている。手法として機械学習によるが、このような前提では、正解（正データ）に対して認識誤り（負データ）の数が極端に少なく、十分な学習効果が得られない。本研究では、重回帰分析、C4.5を基準にクラスの分布に極端に偏りがある訓練事例の場合にも高精度の分類器を構成することを試みる。

2 誤認識文字発見手法の概要

本研究ではOCR認識された文章を入力とし、文字ごとにいくつかのパラメタを抽出した事例データを用い機械学習することにより、誤認識を自動的に検出するシステムを実現する。手法の概要は以下の4つのフェーズからなる。

1. OCR認識文字のパラメタを得るために、統計的確率を算出をする[1]。
2. 訓練データから事例データを作成する。
3. 誤認識文字判定対象であるOCR出力文字のパラメタ抽出をする。
4. 判定モデルを作るための最適な事例データの抽出とAdaBoost[2]を用いた機械学習を行い、判定モデルを導出する。

3 事例データの作成

3.1 パラメタの設定

本研究では、機械学習で用いるパラメタとして以下の8種類を設定する。

[1]衆議院ホームページ[4]より国会会議録約9000万文字のコーパスを得た。

1. 形態素コスト・前後の接続コスト

   単語の構造的正当性の指標として、形態素コストと接続コストを形態素解析プログラム「茶筌」[5]を使用して求めている。形態素解析プログラムは、ある特定の位置からはじまるすべての可能な形態素を辞書引きによって得る。そして、辞書引きによって得られた個々の形態素に対して、その直前の位置に存在するすべての形態素との連接可能性のチェック、および、コストの計算を行なう。

   誤り文字が含まれる形態素は、辞書に該当する形態素が存在しない、または、連接の可能性が低い等が考えられる。つまり、形態素コスト・接続コスト共に他と比較してコストが、高いと推定される。

2. 画数

   複雑な文字ほど、誤認識が起こりやすい可能性がある。そこで、画数データを用いる。各文字に対してそれぞれの画数を割り当てる。

3. 文頭であるか

   パラメタを算出しようとする文字が文頭の場合、前との接続コストを算出することができない。そこで、文頭と文中を区別するパラメタで代用する。文頭である形態素には1を、そうでないときは0を割り当てる。

4. 形態素の長さ

   OCR誤認識文字が含まれる文は、形態素が短く区切られて形態素解析される場合がある。そこで、形態素の長さをパラメタとして用いる。

5. 文字 bi-gram 確率

   OCR誤認識を含む文字列は日本語の文としてに意味を為さない事があると思われる。つまり、OCR誤認識を含む連続した2つの文字の共起確率は低いと考えられる。そこで、本研究ではN-gramモデルを用いる。しかし、tri-gramモデルでは、文字の組合わせが膨大になり、計算量もかかりことから、bi-gramモデルを採用する。

   あらかじめ適切な文字列から構成されたコーパス[4]より、文字bi-gram確率を算出しておく。求めた文字bi-gram確率より、認識文字の文字bi-gram確率を検索し割り当てる。

6. 品詞 tri-gram 確率

   OCR誤認識を含む文字列は日本語的に意味を為さない事がある。つまり、OCR誤認識を含む連続した3つの形態素の品詞共起確率は低いと考えられる。そこで、本研究ではN-gramモデルを用いる。品詞の場合、組合わせ数が少ないためtri-gramモデルを採用する。

   あらかじめ適切な文字列から構成されたコーパス[4]より、品詞tri-gram確率を算出しておく。求めた品詞tri-gram確率より、認識文字の品詞tri-gram確率を検索し割り当てる。

7. 文字誤り易さ確率

   OCR文字認識においては、文字ごとに間違いやすさに偏りがある。この偏りは、大量のOCR誤認識を調査すれば、文字ごとの誤り易さを算出することができる。しかし、手作業で行わなければならず、多大な労力を要する。そこで、同じ文章で画像解像度400dpiの画像と画像解像度200dpiの画像とのOCR結果比較し、差が出た文字を誤りやすい文字とみなす。文字単位で差分が出た文字の数をカウントし差分の出た確率をパラメタとして用いる。本論文では、この確率を文字誤り易さ確率と呼ぶ。

3.2 スムージング

本研究では、N-gramモデルを用いているパラメタがある。しかし、統計的に求めたパラメタにはゼロ頻度問題が存在する。そこで、ゼロ頻度問題に対処するために、スムージングを用いる。

- ワン・カウント法 (One-Count Method)

  本研究では、スムージング手法としてワン・カウント法を用いる。ワン・カウント法は、モデルの次数、学習データ量にかかわりなく優れた性能を示すとされている[1]。

4 分類器の生成

OCR誤認識文字の検出では、OCR出力を「正」と「負」の2クラスに分類する。ここでは、線形モデルを作ることのできる重回帰分析を用いる。

4.1 ブースティング

本研究で使用する訓練データは、クラスの数に極端な偏りがある。弱判定器にしきい値の調整が容易な重回帰分析を用いたAdaBoostアルゴリズム[2]により高精度の分類器を作成する。$D_t(i)$はラウンドtで求められた事例iの重みである。

各ラウンドtの弱判定器は、事例ベクトル$(X_{i1}, X_{i2}, \cdots, X_{im-1})$とその重みの積

$$D_t(i) \cdot \hat{Y}_i = D_t(i) \cdot (b_0 + \sum_{j=1}^{m-1} b_j X_{ij}) \quad (1)$$

と、判定クラスの重みの積

$$D_t(i) \cdot Y_i \quad (2)$$

から残差に関する最小2乗法により最適な $b_0, b_1, \cdots, b_m$ を求める。

4.2 事例データの削除

使用する事例データは、クラスの分布に極端な偏りがあり、的確なブースティングを行うのは困難である。そこで、事例データ中の割合が多い正データに限り削除する。削除は、訓練データとして用いるデータをクローズドな環境で重回帰分析を行い、正解を誤判定した事例を削除し、正データ数を減らす。本研究ではこの手法により、正データを10分の1まで削減した。

4.3 再現率・適合率

本研究はOCR誤認識文字を検出することを主眼を置いているため、「負」クラスのみに着目した再現率・適合率・F値を求める。

以下のように定義した再現率、適合率を求め、検討を行なう。本研究では、OCR誤認識文字検出を目的としているため、誤認識文字をどれだけ検出できたかを表す再現率を重視する。このため $\gamma = 3$ とする。

$$\text{再現率} R = \frac{\text{「誤」判定が正しかった文字数}}{\text{実験データ中の誤り文字数}} \quad (3)$$

$$\text{適合率} P = \frac{\text{「誤」判定が正しかった文字数}}{\text{「誤」と判定した文字数}} \quad (4)$$

$$F = \frac{(1+\gamma^2) \times P \times R}{\gamma^2 \times P + R} \quad (5)$$

5 実験結果と評価

5.1 実験データについて

誤認識文字判定モジュールで行う訓練と評価のために、国会議事録55536文字からなるOCR認識文字の特徴ベクトルを抽出し、正誤のクラスに分類したものを作成する。「正」、「負」のクラスの分類に際しては、OCR認識対象文書となった国会会議録原本のコピーを参照し、手作業で行う。目標とする99.95%を達成するための再現率 R を以下の式に示し、表1に訓練データの内訳を示す。なお、OCR出力の文字認識制度を f で表す。

$$\text{目標を達成するための再現率} R = \frac{99.95 - f}{1 - f} \quad (6)$$

よって、本研究で目標とする99.95%の精度を実現するには、76.37%以上の再現率が必要となる。

総文字数	正	誤	OCR認識精度
55536	55419	117	99.789%

表 1: 実験データの内訳

5.2 実験結果

実験結果と評価、検討を示す。

1. オリジナルデータを用いた重回帰分析の結果

   OCR出力から事例データを作成し、そのデータをそのまま用いて機械学習を行った結果を2に示す。

No.	R	P	F
1	0.452	0.072	0.30%
2	0.667	0.092	0.41%
3	0.528	0.207	0.46%

表 2: オリジナルデータを用いた重回帰分析の結果

2. 重回帰分析を用いて正データを削除したデータを用いた結果

   クローズドな環境で重回帰分析を行い、そこでの誤判定データを削除したデータを用いた重回帰分析の結果を表3に示す。

No.	R	P	F
1	0.484	0.143	0.39%
2	0.576	0.373	0.55%
3	0.811	0.160	0.58%

表 3: 重回帰分析を用いて正データを削除したデータを用いた結果

3. ブースティングを行ったデータを用いた結果

   クローズドな環境で誤判定データを削除したデータに、ブースティングを行ったデータを用いた重回帰分析の結果を表4に示す。

No.	R	P	F
1	0.355	0.224	0.34%
2	0.848	0.222	0.66%
3	0.811	0.167	0.59%

表 4: ブースティングを行ったデータを用いた結果

4. ランダムに正データを削除したものを用いた結果

   比較対象として、ランダムに正データを削除したものを用いた重回帰分析の結果を表 5に示す。

No.	R	P	F
1	0.452	0.126	0.36%
2	0.667	0.112	0.45%
3	0.623	0.213	0.52%

表 5: ランダムに正データを削除したものを用いた結果

5. C4.5 を用いた結果

   比較対象として、分類器に C4.5 を用いた結果を表 6に示す。

No.	R	P	F
1	0.642	0.290	0.57%
2	0.612	0.578	0.61%
3	0.500	0.226	0.45%

表 6: C4.5 を用いた結果

ブースティングを用いた手法では、表 4で示すように、No.1 以外のテストデータに関して、他の手法よりも良好な結果が得られた。しかし、No.1 のテストデータは、重回分析を用いて正データを削除したデータを用いた結果よりも、ブースティングを行った結果悪化した。これは、ブースティングの特性として、基本的に弱分類器には 50% 以上の分類精度が求められているからであると考えられる。

6 おわりに

本研究では、文字ごとに文の構造、意味、OCR システムによる誤認識文字の偏りに関する特徴ベクトルを抽出し、OCR 出力から事例データを作成した。次に、弱判定器として重回帰分析を用いた AdaBoost アルゴリズムにより、クラスの分布に極端に偏りがある訓練事例の場合にも高精度の分類器を構成することを試みた。

本研究では、人手による後処理の補助となるようなシステムの構築を目標としている。方針として、OCR 誤認識文字をできるだけ検出することを主眼を置き、OCR 誤認識文字を検出するものである。OCR 誤認識文字の検出には重回帰分析と機械学習システム C4.5 を用い、OCR 誤認識文字を検出することで実現した。

機械学習を行う際、形態素コスト、接続コスト、文字誤り易さ確率等、合計 8 種のパラメタを用いた。また、3 分割交差検定によりシステムの評価を行った。訓練データに関しては、現在の OCR システムに合致させるために、実際の OCR 出力を用意した。評価においては、OCR 誤認識文字をできるだけ検出することを主眼を置いているため、再現率を重視した。その結果、AdaBoost アルゴリズムを用いた機械学習では目標としている OCR 誤認識文字を訂正した後の誤り文字含有率 99.95% を達成した。

参考文献

[1] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, Morgan Kaufmann Publishers*, pp. 310–318, 1996.

[2] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pp. 23–37, 1995.

[3] 久光徹, 丸川勝美, 島好博, 藤澤浩道, 新田義彦. ocr 誤認識後処理の効率について. 情報処理学会自然言語処理研究会研究報告 NL-104, pp. 17–24, 1994.

[4] 衆議院. http://www.shugiin.go.jp.

[5] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム「茶筌」version2.0 使用説明書 第二版. NAIST Technical Report, 奈良先端科学技術大学院大学 松本研究室, 1999.

[6] 竹内孔一, 松本祐治. 共起情報と統計的形態素解析による ocr 誤り訂正. 情報処理学会自然言語処理研究会研究報告 NL-121, pp. 17–24, 1997.