tri-gram を用いたOCR誤り検出における新聞記事、社説、コラム別の結果の比較

松山 高明　　渥美 清隆　　増山 繁
matsu@smlab.tutkie.tut.ac.jp　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

紙に印刷された形態の文書を機械可読データに変換する手段としてOCRが利用されるが、変換後の文書中には通常は誤認識された文字が多数含まれる。この誤認識文字の検出に関する様々な手法が提案されている。その中には解析処理の失敗により誤りを検出する形態素解析による手法や、*tri*-gram 等、言語情報を統計的に扱うことで誤りを検出する手法などがある。近年、大量の機械可読なコーパスが利用可能になったため、統計的手法である *tri*-gram の研究が盛んに行なわれている [1], [2]。本稿ではそのような *tri*-gram に着目し、OCR の誤り検出能力の検討を行なう。

本研究では従来の研究に習い、誤り検出能力を評価するために適合率と再現率を評価基準として用いる。文献 [3] では *bi*-gram、*tri*-gram、4-gram について述べた。しかし、*bi*-gram は文字の共起する確率が高くなり過ぎ、良好な再現率を得ることはできない。また、4-gram では逆に未知文字列が増え過ぎ、適合率が下がる。ある程度の適合率を得るためには、大量のコーパスを用意すればいいが、現状では十分なコーパス量を用意できない。そこで本稿では *tri*-gram に着目する。本稿では新聞記事、社説、コラム別に関しての検出能力の違いを比較し、検討を行なう。

2 誤り検出方法

2.1 *tri*-gram の定義

文字 a_i の出現確率がそれより二つ前までの文字列 $a_{i-2}a_{i-1}$ の出現により決定される確率モデルを *tri*-gram と呼ぶ。文字 a_i の出現確率は次の条件付き確率で計算する。

$$p(a_i|a_{i-2}a_{i-1}) = \frac{p(a_{i-2}a_{i-1}a_i)}{p(a_{i-2}a_{i-1})} \qquad (1)$$

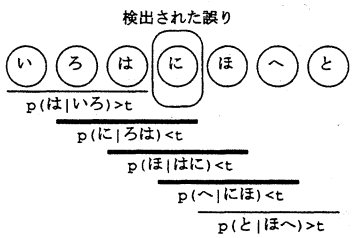

図 1: 誤り検出例

$p(a_{i-2}a_{i-1})$　：$a_{i-2}a_{i-1}$ の出現する確率
$p(a_{i-2}a_{i-1}a_i)$　：$a_{i-2}a_{i-1}a_i$ の出現する確率
$p(a_i|a_{i-2}a_{i-1})$　：$a_{i-2}a_{i-1}$ の次に a_i の出現する確率

2.2 誤り検出法

2.1の手法でコーパスから文字の出現確率を求め、確率連鎖辞書を作成する。

誤りの含まれた認識文書の誤り文字の検出は認識文書中の文字列を確率連鎖辞書を用いて出現確率を調べ、その値があらかじめ設定した足切り値よりも小さい文字列が3つ重なる文字を誤り文字と推定する。図1は文字「に」を誤り文字として検出した例である。t は足切り値である。

2.3 実験対象

日本経済新聞 CD-ROM 版 (1990年 ～ 1995年) を用いて実験を行なう。これより実験対象として「社説」、コラム「春秋」、コラム「アーバンNOW」、さらにキーワード「政治」を持つ記事からコラムを除いた記事を実験対象として選んで用いる。

表 1: OCR 認識対象の作成

認識対象	プリンタ	認識率
「社説」	EPSON LP-1600	91.6 %
「春秋」		92.0 %
「アーバンNOW」		90.2 %
「政治」	NEC PC-PR3000PS/4	92.7 %

表 2: 実験に用いるコーパス

	変化幅	総コーパス量
「社説」	約 160kbyte	約 10Mbyte
「春秋」	約 50kbyte	約 3Mbyte
「アーバンNOW」	約 90kbyte	約 5.5Mbyte
「政治」	約 150kbyte	約 8.7Mbyte

2.4 OCR 認識文書の作成

OCR 認識文書として選んだテキストを表 1 [1]で示すプリンタにより印刷する。スキャナ[2]で対象テキストを画像として取り込み、OCR ソフト[3]を用いて OCR 認識文書を作成する。

2.5 評価基準

本研究でも従来の研究に習い、誤り検出能力を評価するために適合率と再現率という評価基準を採用する。なお、適合率・再現率は次式で定義される。

$$\text{適合率} = \frac{\text{正しく誤り推定できた文字数}}{\text{誤りと推定した文字数}} \times 100[\%] \quad (2)$$

$$\text{再現率} = \frac{\text{正しく誤り推定できた文字数}}{\text{誤り文字数}} \times 100[\%] \quad (3)$$

適合率とは誤り文字をより的確に指摘した割合であり、再現率とは誤り文字を漏れなく指摘した割合である。これら両方の割合が高いほど誤り検出能力が高い。

3 実験

3.1 実験方法

実験対象として選んだ「社説」、「春秋」、「アーバンNOW」、「政治」からそれぞれ 10 編を OCR 認識対象として選択する。適合率・再現率はこの 10 編の平均を求め、誤り検出能力の比較を行なう。ここで 10 編に選ばなかった残りのコーパスを確率連鎖辞書作成用のコーパスとして使用する。各々の実験対象に対し、コーパス量を次第に増加させる実験を行なう。また、コーパスの総量と変化幅は表 2の通りである。なお、変化幅はコーパス量の 1 年分の $\frac{1}{10}$ で、総コーパス量は 6 年分である。

なお、足切り値は 0.001 を採用する。

[1]都合によりプリンタを変更した。
[2]EPSON GT-6500(300dpi)
[3]バーズ情報科学研究所「The OCR 日・英」

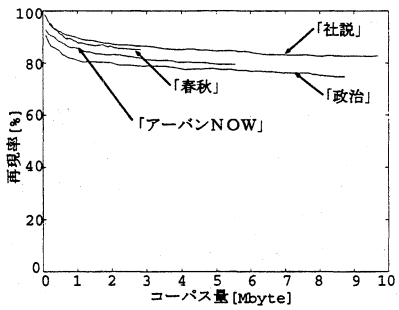

図 2: 再現率

3.2 実験結果

図 2に再現率、図 3に適合率の結果を示す。

図 2の再現率の結果は「社説」と「春秋」でほとんど同程度の近い値を示しているのに対し、「アーバンNOW」と「政治」は少し低めである。しかし、どれも 80% 前後の値であり比較的高い値である。

図 3の適合率の結果は「春秋」、「アーバンNOW」、「政治」がほぼ同程度の値を示しているのに対し、「社説」だけはそれらよりも高い値を示している。

3.3 再現率の差に関する考察

再現率に差が出た理由を考察する。誤り推定し落とした誤り文字には、どのような文字があるかを調べる。すると平仮名やカタカナが誤り文字である場合に誤り推定し落とすことが多いことがわかった。

再現率の低い誤り検出対象の認識文書を見てみると、「ニュース」という言葉が「ニユース」となっていたり、「でしょう」が「でしよう」になっている箇

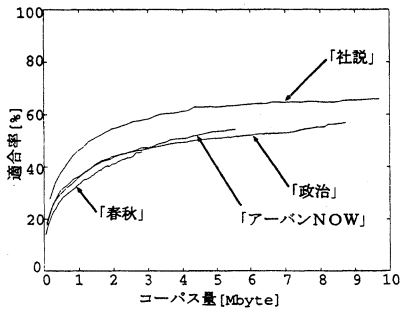

図 3: 適合率

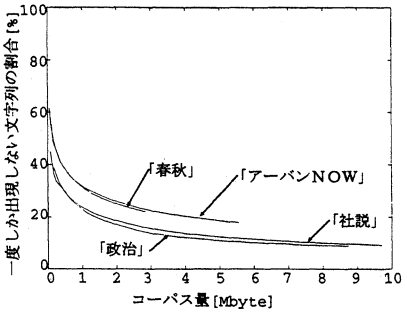

図 4: 一度しか出現しない文字列の割合

所が多く見られた。このように小さい仮名文字が大きい仮名文字に認識されて出力されるという現象はたびたび起こることである。「ニユース」という言葉の様に「ユース」という言葉があるために誤り推定できないでいるのである。しかもこの「ニユース」と誤認識されている場合には他に出現する「ニュース」も同様に「ニユース」と誤認識され、推定し落としてしまっているのである。このような単語が多く出現すると再現率を低下させる原因となる。

なお、「政治」での誤り推定し落とした誤り文字には「だった。」が「だったり」と「。」が「り」に誤認識されているケースが多くあった。こういった誤り文字が誤り推定できないため、再現率が他よりも低めの値を示している。

3.4 適合率の差に関する考察

同一のコーパス量での検出能力は、図2で示すように再現率はそれぞれが近い値を示したのに対し、図3が示す適合率の結果は「社説」のみが高い値を示しており、その他はほとんど同程度の値を示している。文献[3]では適合率の値に差が出るのはコーパスの種類により一度しか出現しない文字列の割合が違うためだと述べた。一度しか出現しない文字列の割合を使うことによってコーパスの種類による適合率の違いを近似できる。

次式により一度しか出現しない文字列の割合を求める。なお、本稿では *tri*-gram を扱っているので一文字列長は 3 文字である。

$$\text{一度しか出現しない文字列の割合} = \frac{\text{一度しか出現しない文字列の個数}}{\text{文字列の述べ語数}} \times 100[\%] \quad (4)$$

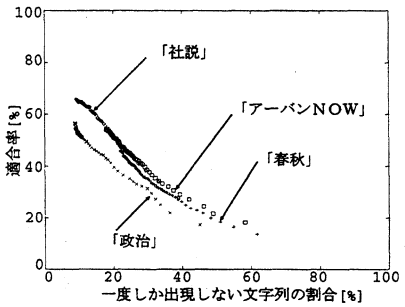

図 5: 適合率の近似

一度しか出現しない文字列の割合は図4のようになった。「春秋」と「アーバンＮＯＷ」が同程度、「社説」と「政治」が同程度の割合を示している。「社説」と「政治」の適合率は大きく違うにも関わらず、一度しか出現しない文字列の割合は同程度となった。

さらに、適合率と一度しか出現しない文字列の割合の関係を図5に示す。「社説」と「春秋」と「アーバンＮＯＷ」は同一の一度しか出現しない文字列の割合で同程度の適合率を示している。一方、「政治」に関してはこの近似には当てはまらなかった。これについて考察する。

「政治」で使用したコーパスは純粋な政治記事だけではなかった。キーワード「政治」がふられている記事には純粋な政治記事だけではなく、文章内に政治という言葉があればキーワードに「政治」がふられている。それらの記事はコラムなら取り除くことが可能だが、それ以外の記事については取り除くことができないでいる。また、ここで対象として適切な記事は文章

のみからなる記事であるが、「政治」に関する記事なのでアンケートの結果や人物の名前や日程などの文字列の羅列もコーパスに含まれる。さらに、文章中に「政治」という単語があるだけで政治とはかけ離れた内容の記事もある。一度しか出現しない文字列の割合による近似に合わないのはこれらの混入によるものと考えられる。つまり、人の名前などの文字列の羅列は出現頻度が低いにも関わらず、二度以上出現している場合が多い。そのため、一度しか出現しない文字列の割合には含まれないので、一度しか出現しない文字列の割合は低くなる。

このような記事の含まれたコーパスを用いて確率連鎖辞書を作成しているために、「コラムなどのテーマごとに一度しか出現しない文字列の割合が変わり、その差が適合率の値の差として現れる」という提案には当てはまらなかったのだと考えている。

4 むすび

本稿で行なった実験により再現率は同一の実験条件の下ではコーパスの種類に関わりなく同一のコーパス量で同程度の値を示すことが分かった。

適合率はコーパス量の代わりに一度しか出現しない文字列の割合を用いて近似することを試みたが、うまくいかなかった。これは「政治」というキーワードのふられた記事の摘出法に問題があるためだと考えられる。

今後の課題としてはキーワード検索の際に純粋な政治関連記事だけを取り出して、統一された条件下での実験を行ない、その上で一度しか出現しない文字列の割合によって適合率を近似できるかについての検証があげられる。

参考文献

[1] 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」, 情処研究報告, NL-102-13(1994)

[2] 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」, 情処研究報告, NL-102-12(1994)

[3] 松山, 渥美, 増山:「n-gramによるOCR誤り検出の能力検討のための適合率と再現率の推定に関する実験と考察」, 言語処理学会 第2回年次大会,129(1996)

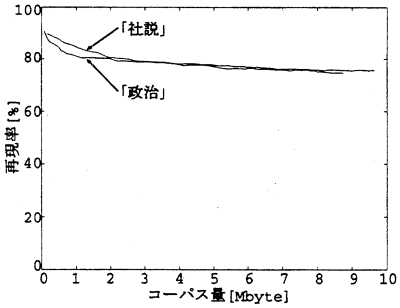

図 6: 同一プリンタでの再現率

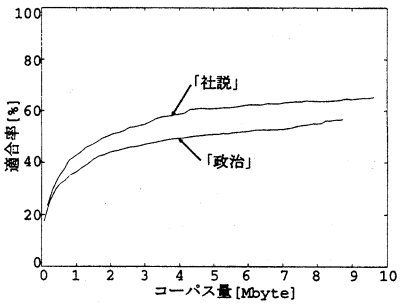

図 7: 同一プリンタでの適合率

A 付録：同一プリンタで比較

表1で示すように「政治」では他とはプリンタを変更して実験している。そこで「社説」もNECのプリンタで印刷して実験を行ない、「政治」との差異を表した結果を図6、図7に示す。なお、NEC PC-PR3000/PS4で印刷してOCR認識文書を作成した場合の「社説」の認識率は、93.9%であった。

図6が示すように再現率については同一コーパス量での「社説」と「政治」の能力値がほとんど一致したのに対し、図7が示すように適合率については能力差はほとんど縮まらなかった。