OCRの誤り検出における2重マルコフ連鎖の一般的な能力の検討

4R－6

松山 高明　渥美 清隆　増山 繁
豊橋技術科学大学 知識情報工学系

1 はじめに

文書の情報を，紙による印刷された形態ではなく機械可読データとして提供することによって，データの編集や検索も便利になる．紙に印刷された文書を機械可読データに変換する方法として，OCRを利用する方法がある．しかし，誤認識された文字の修正に手間を割かねばならない．

OCRの誤り検出の研究は従来精力的に行なわれているが[1], [2]，これらの研究の多くは実験に用いる文章の種類を制限しているので，一般的な誤り検出能力が検証できているわけではない．

本研究では実際にOCRで取り込んだ文章に対し2重マルコフ連鎖を用いた誤り検出を行なう実験を行ない，誤り検出に対する2重マルコフモデルの一般的な性質を検討する．

2 2重マルコフモデルの説明

2重マルコフモデルを用いてどのように誤り検出するかを述べる．簡単にいえば，コーパスから文字列の連なる確率の計算結果を学習辞書として持ち，入力文字列における文字列の連なる確率の低い部分を発見すれば良い．計算手順は次の通りである．

学習文字列を $w_1, w_2, ..., w_n$ とする．各々の文字についての条件つき確率 $p(w_{i+2}|w_i w_{i+1})$ を計算し，2重マルコフモデル確率連鎖辞書を作る．

入力文字列を $v_1, v_2, ..., v_n$ とする．ここから3文字列 $v_i v_{i+1} v_{i+2}$ を取り出し，先ほど作った確率連鎖辞書から $p(v_{i+2}|v_i v_{i+1})$ を求める．そして，足切り値 T を定めておき，$T \geq p(v_{i+2}|v_i v_{i+1})$ が3回以上連続して起こるとき，誤り文字列を推定できる．

3 実験材料

日本経済新聞の1990年，1992年の記事を収録したCD-ROM中から，比較的統一された文体で書かれている「社説」と形式的でない口語調でかかれている「春秋」を選び，これらの記事を使用する．

600dpiのレーザープリンタでそれぞれの記事を印刷し，EPSON GT-6500(300dpi) とBIRDS社製のThe OCR 日・英 Ver.1.0 を用いてOCR認識文書を作成する．認識率は91.8%～95.3%で平均94.3%であった．

4 実験内容

- 「社説」，「春秋」で学習量は1年分固定で足切り値を変化させて実験する．
- 「社説」，「春秋」で足切り値は0.001に固定して学習量を変化させて実験する．
- 「春秋」で大きなコーパスを用い学習量を変化させて実験する．

5 実験結果

評価基準として以下の適合率と再現率を採用する．

$$\text{適合率} P = \frac{\text{正しく誤りを推定した文字数}}{\text{誤りを推定した文字数}}$$

$$\text{再現率} R = \frac{\text{正しく誤りを推定した文字数}}{\text{実際に誤りである文字数}}$$

ここで示す適合率・再現率のグラフはすべて，10記事についての平均である．

5.1 足切り値の変化について

図1は「社説」で足切り値を0.0001 ～ 0.0009, 0.001～0.009と変化させたグラフを示す．また，図2は「春秋」で，0.001～0.009と変化させたグラフを示す．

両グラフとも再現率の方は，値が大きくなるに従って単調に増加し，適合率は単調に減少している．

全体的に，適合率は低く再現率は高いので，適合率を優先して考えると，0.001付近の足切り値を採用するのが良いと考えることができる．

5.2 学習量の変化について

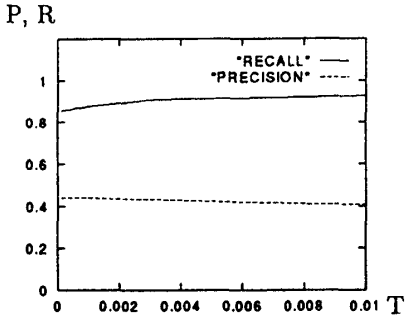
図3に「社説」で学習量を変化させた時のグラフを示す．足切り値は0.001に固定して行った．

90年度の1年分のコーパスで学習量を変えて実験を行い，さらに92年度のコーパスからOCRに用いた記事を除いた1年分を追加して実験を行った．コーパスの大きさは2年分で3MByteあった．

利用可能な「社説」の記事がこれだけだったので，これ以上の学習をすることはできなかったが，図3のグラフより適合率は上昇するが，再現率は低下するという予

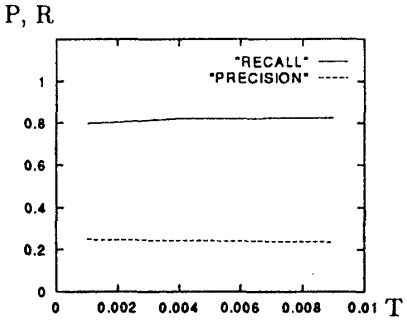
†Experimental Results on OCR Error Detection Capability Using Second Order Markov Model

‡Takaaki Matsuyama, Kiyotaka Atsumi, Shigeru Masuyama, Knowlege-based Information Engineering , Toyohashi University of Technology



(P: 適合率, R: 再現率, T: 足切り値)

図 1: 社説で足切り値を変化させたとき



(P: 適合率, R: 再現率, T: 足切り値)

図 2: 「春秋」で足切り値を変化させたとき

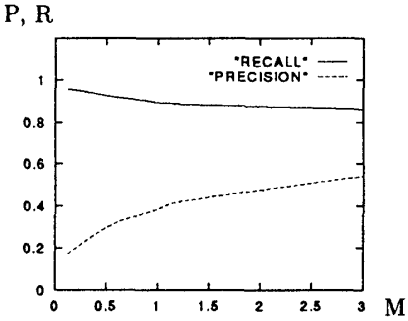
測が立てられる. 適合率は, 0.5 程度は最低限必要と思われるので, 2.5MB 程度の学習量が良いと考えられる.

次に, 「春秋」を使って実験を行なった. この結果を図 4に示す. コーパスは同じく 2 年分であったが大きさは 0.93MByte だった. これを用いて, 適合率は 0.25 程度にしかならなかった. 「社説」の場合, 1MByte の学習量で適合率が 0.35 であったことを考えるとかなり低いといえる.

5.3 大量の学習を行う

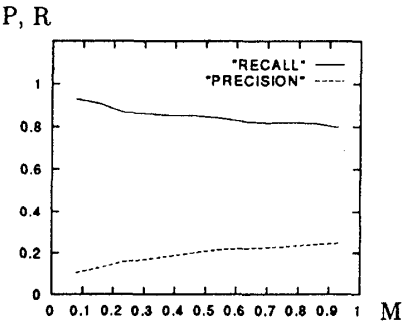
適合率の低い「春秋」も学習量を増やせば適合率を上昇させることができるのではないかと考え, 学習量を大幅に増やして実験を行なってみた. ただし, 学習用コーパスは「春秋」に限定せず, 他の日経新聞のコラム (「社説」「文化」「アーバン・ナウ」「婦人」) も使用した. この結果を図 5に示す.

このグラフを見る限り, 10MByte のコーパスを用いても目安としている適合率 0.5 には達しなかった. これ以上学習量を増やしても再現率が下がってくるので「春秋」のような形式的でない口語調のテキストには 2 重マルコフモデルによる誤り検出は適さないのではないかと考えられる.



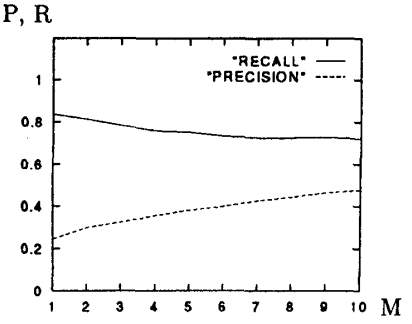
(P: 適合率, R: 再現率, M: 学習量 [MByte])

図 3: 社説で学習量を変化させたとき



(P: 適合率, R: 再現率, M: 学習量 [MByte])

図 4: 「春秋」で学習量を変化させたとき



(P: 適合率, R: 再現率, M: 学習量 [MByte])

図 5: 「春秋」で大量の学習量を変化させたとき

6 まとめ

以上の結果より, 足切り値は 0.001 に設定し, 学習用コーパスを 2.5MByte 程度にすると, 「社説」のような統一された文体に対し, ある程度の性能で誤り検出できることがわかった. また, 「春秋」のような口語調の文体には, 学習量を上げてもそれほど良い性能は達成できないことも分かった.

参考文献

[1] 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語処理」, NL 102-12, 情報処理学会 (1994).

[2] 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」, NL 102-13, 情報処理学会 (1994).