# 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

松山 高明　matsu@smlab.tutkie.tut.ac.jp
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

本研究では、実際にOCRで取り込んだ文章に対しn-gramを用いた誤り検出を行なう実験を行ない、n-gramの一般的な誤り検出能力を検討している[3]。

本稿では、新聞記事の特徴と適合率の関係を調べるために、学習コーパス中で一度しか出現していない文字列の割合を調べた。そして、この関係を回帰方程式で近似するとうまく表せることが分かった。この関係についてのn-gramの一般的な誤り検出能力の検討を行なう。

## 2 誤り検出実験

### 2.1 実験方法

実験は全て、文字単位で行う。

例として、tri-gram を用いてどのように誤り検出するかを述べる。つまり、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 $w_1w_2...w_n$ とする。各々の文字についての条件つき確率 $p(w_{i+2}|w_iw_{i+1})$ を計算し、確率連鎖辞書を作る。

入力文字列を $v_1v_2...v_n$ とする。ここから3文字列 $v_iv_{i+1}v_{i+2}$ を取り出し、先ほど作った確率連鎖辞書から $p(v_{i+2}|v_iv_{i+1})$ を求める。

そして、足切り値 $T$ を定めておき、

$$p(v_{i+2}|v_iv_{i+1}), p(v_{i+3}|v_{i+1}v_{i+2}), p(v_{i+4}|v_{i+2}v_{i+3}) \leq T$$

となるとき、$v_{i+2}$ を誤り文字列と推定する。

### 2.2 実験対象

日本経済新聞の1990年の記事を収録したCD-ROMと1992年の記事を収録したCD-ROMを用いる。

その中から、「社説」と一面に掲載される「春秋」というコラムを選んだ。「社説」は比較的統一された文体で、「春秋」は口語調で書かれている。

誤り検出対象は、「社説」「春秋」、それぞれ10記事をランダムに選び出す。さらにそれを、600dpiのレーザープリンタでそれぞれの記事を印刷し、スキャナ*とOCRソフト†を用いてOCR認識文書‡を作成する。

学習対象は、「社説」「春秋」の記事、それぞれ2年分から誤り検出対象とする記事を除く。それを10等分し、学習量を変化させる実験に用いる。

### 2.3 評価方法

評価基準として以下の2つを採用する。

$$\text{適合率} = \frac{\text{うまく誤り指摘できた文字数}}{\text{全誤り指摘文字数}} \times 100[\%]$$

$$\text{再現率} = \frac{\text{うまく誤り指摘できた文字数}}{\text{全誤り文字数}} \times 100[\%]$$

## 3 これまでの実験経緯

文献[3]では、次のようなtri-gramによるOCR誤り検出について実験し検討をした。

- 足切り値を変化させて実験
- 学習量を変化させて実験
- 学習対象と誤り検出対象を同じ種類に限らず、学習を大量に行って実験

これにより、足切り値は0.001を使うのが最も良いということが分かった。また、「社説」と「春秋」では適合率・再現率ともに「社説」の方が良い結果を示し、これは「春秋」では一度しか使われていない文字列が多いために、学習した辞書中にその文字列が存在しないことが多いからではないかと考察した。

## 4 一度しか出現しない文字列の割合と適合率、再現率の関係

### 4.1 一度しか出現しない文字列の調査

前節で述べたような考察を確かめるために、一度しか出現しない文字列の割合がどのくらいかを次の式に基づき調査する。

$$\frac{\text{一回しか出現しない文字列の個数}}{\text{文字列の述べ語数}} \times 100[\%]$$

*EPSON GT-6500(300dpi)

†BIRDS社製の The OCR 日・英 Ver.1.0

‡認識率は91.8% ～ 94.3% であった

また、一度しか出現しない文字列の割合がより多いと思われる 4-gram と、一度しか出現しない文字列の割合がより少ないと思われる bi-gram についても実験し、比較検討する。

「社説」「春秋」それぞれについて、学習量を変化させて実験した時と同じ量のコーパスで、一度しか出現していない文字列の割合を調査する。

そのような調査をした結果、図 2の様になった。「春秋」ではその割合が高く、また、4-gram での割合も高い。

この結果を見てみると、一度きりしか使われていない文字列の割合が高いもの（特に 4-gram、「春秋」）では、未知文字列の比率が高くなり、適合率が極端に低くなる。

## 4.2 適合率、再現率との関係

ここで述べる適合率や再現率はすべて足切り値を 0.001 にした時の値である。

### 4.2.1 適合率

「一度しか使われていない文字列の割合」と「適合率」の関係について調べた。以下は、10 記事の適合率の平均である。

$x$ は「一度しか出現しない文字列」であり、$y$ は「適合率」である。また、$\rho$ は相関係数である。これをグラフにプロットすると、うまく分数関数で表されそうな形になった。

そこで、（厳密には分数関数ではないが）次のように回帰方程式を利用して近似することにした。

1. 適合率を $y$、一度だけしか出現しない文字列の割合を $x$ とする
2. $Y = y, X = \frac{1}{x}$ と変数変換する
3. Y の X についての回帰方程式を求める
4. 変数を戻してできた回帰方程式を求める

結果は次のようになった。

$$\text{bi-gram: } y = \frac{109}{x} - 33.5,\ \rho = 0.931$$

$$\text{tri-gram: } y = \frac{929}{x} - 3.76,\ \rho = 0.995$$

$$\text{4-gram : } y = \frac{1508}{x} - 13.6,\ \rho = 0.998$$

これだと、平均した値を用いているために、偶然この曲線に近くなっただけとも考えられ、正確に表せているとは断言できない。そこで、次のように平均ではなくて元のデータで計算を試みた。

図 3、図 4、図 5は、10 記事を個々のデータとしてプロットして回帰方程式を求めたグラフである。

$$\text{bi-gram : } y = \frac{109}{x} + 33.5,\ \rho = 0.827$$

$$\text{tri-gram: } y = \frac{929}{x} - 3.76,\ \rho = 0.916$$

$$\text{4-gram : } y = \frac{1508}{x} - 13.6,\ \rho = 0.886$$

当然、相関係数は落ちているものの、それでも $\rho = 0.9$ 前後という高い値を保てている。これは、それぞれのデータのバラつきが小さかったためであろう。

### 4.2.2 再現率

同じように、再現率についても計算を試みた。これについては、そのまま回帰方程式を適用した。

10 記事の再現率の平均との関係の回帰方程式は次の通りである。

$$\text{bi-gram : } y = -0.211x + 68.1,\ \rho = -0.203$$

$$\text{tri-gram: } y = 0.0824x + 85.0,\ \rho = 0.334$$

$$\text{4-gram : } y = 0.0425x + 92.4,\ \rho = 0.534$$

図 6、図 7、図 8に再現率の元のデータでのグラフを示す。回帰方程式は次の通りである。

$$\text{bi-gram : } y = -0.211x + 68.1,\ \rho = -0.127$$

$$\text{tri-gram: } y = 0.0824x + 85.0,\ \rho = 0.155$$

$$\text{4-gram : } y = 0.0425x + 92.4,\ \rho = 0.101$$

図のように、再現率については、あまり良い結果は出なかった。

bi-gram では、本来右上がりになるべき再現率のグラフが右下がりになってしまっている。これは、点があまりにも左に寄ってしまっているために、回帰方程式による近似では少し無理があったといえる。

## 4.3 グラフの形についての考察

学習量を増加させる、つまり学習コーパスの量を増加させると、一度しか現れない文字列の割合は減少する。グラフでは横軸で右に行くほど学習コーパスの量が少なくなっている。

図 1に一度しか現れない文字列の割合が増加した時の極端な例を示す。

一度しか現れない文字列の割合が増加すると、検索する文字列が辞書中に存在せずに、誤りと推定してしまう確率が増えるので、誤り推定範囲が大きくなる。すると、全誤り指摘文字数も増加し、うまく誤り指摘できた文字数も増加する。しかし、誤り文字よりも正しい文字の方が数が多い。つまり、一度しか現れない文字列の割合が増加すると、適合率の式の分母は早く増加し、分子はゆっくり増加するので適合率は下がるのである。図 1では $\frac{2}{6} \rightarrow \frac{3}{11}$ となる。

一方、全誤り文字数は一定なので、一度しか現れない文字列の割合が増加すると、再現率は増加するのである。図 1では $\frac{2}{4} \rightarrow \frac{3}{4}$ となる。

○：正しい文字
×：誤り文字

○（○×○×○○）○×○○○○×○
↓一度しか…の割合が増加
（○○×○×○○○×○○）○○×○

図 1: 一度しか現れない文字列の割合が増加した時の例

### 4.4 bi-gram, tri-gram, 4-gram についての考察

回帰方程式によって近似した適合率のグラフが、tri-gram と 4-gram で、どうして形に違いが出たのかを考察する。また、ここで述べることは、そのまま bi-gram と tri-gram のグラフの違いについても言える。

一回しか出現しない文字列の割合が小さい時について考える。tri-gram は、4-gram よりも 1 文字分だけ短い文字列を扱っているので、一回しか出現しない文字列の割合が同じならば、tri-gram の方が、偶然一致する可能性が高くなるのである。つまり、tri-gram では、誤り文字を指摘し落としてしまうので、4-gram の方が適合率が高くなるのである。

次に、一回しか出現しない文字列の割合が大きい場合について考える。この割合が大きいという事は、学習コーパスが小さく、辞書にある文字列の種類も少ないので、正しい文字でさえも誤り推定してしまう可能性が高くなる。つまり、tri-gram の方が 1 文字分だけ短い文字列を扱っている分、文字列が辞書中に見つかる可能性が高くなり、誤り文字と推定する文字が減るので、tri-gram の方が適合率が高くなるのである。もちろん、正しい文字だけでなく、誤り文字も落としてしまう可能性も高くなるが、誤り文字よりも正しい文字の方が絶対的に多いので、この事が言える。

bi-gram は適合率は高いものの再現率が低く、検出し落す誤り文字が多いので、妥当な手法とはいえないだろう。また、4-gram は学習量が十分ある時には良い結果を示すが、十分ない時にはかなり低い適合率になる。通常はそんなに多くの学習辞書を用意できないので、tri-gram を用いるのがもっとも適した方法といえるだろう。

### 4.5 再現率と適合率についての考察

一回しか出現しない文字列と適合率の関係をグラフに表して、回帰方程式にあてはめると、相関係数がかなり高い値が出た。それに対し、一回しか出現しない文字列と再現率の関係を、同じようにグラフに表して回帰方程式にあてはめてみたところ、これはあまり良い結果にはならなかった。これの理由を考察する。

適合率では、「全誤り指摘文字数」が増えれば増えるほど、「うまく誤り指摘した文字数」も、増加する。つまり、適合率の場合は、分子が増えれば、分母が増え、分子が減れば分母も減る。これにより、お互いの変動が打ち消し合って、求まる適合率は、バラつきが小さくなり、うまく回帰方程式に乗ると考えられる。

次に、再現率の「全誤り文字数」は、学習量や足切り値を変化させても常に一定で、対象とする記事を変えない限り、常に同じ文字数である。これにより、「うまく指摘できた文字数」のバラつきがそのまま再現率の値に出てしまうために、回帰方程式上にはうまく乗らなかったのだと考えられる。

上記の適合率に関する考察を確かめるため、「うまく誤り指摘できた文字数」と「誤り指摘文字数」の平均の値からの差をプロットしグラフにして回帰方程式を計算した。

「社説」「春秋」の tri-gram での結果に適用しすると、

$$「社説」：y = 0.314x + 0.168(\rho = 0.904)$$

$$「春秋」：y = 0.284x + 0.196(\rho = 0.754)$$

となった。この結果を図 9、図 10に示す。

この結果が示すように、「誤り指摘文字数」が増えると、それにともなって「うまく誤り指摘できた文字数」も増えている。つまり、適合率は分子と分母で、お互いの変動をうち消しあって、適合率そのもののバラつきが小さくなっているといえる。

## 5 今後の課題

学習コーパス中で一度しか出現しない文字列の割合から適合率を推定する本手法は、学習する確率連鎖辞書の元コーパスを一度用意してから計算しないといけないという点が不便である。誤り検出の対象とするテキスト文書から、どのくらいの適合率や再現率を推定することができるのか割出すための手法についても検討したい。

## 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後処理」, 情処研究報告, NL-102-12(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」, 情処研究報告, NL-102-13(1994)

[3] 松山, 渥美, 増山: 「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」, 情報処理学会 第 51 回全国大会, 2-175(1995)


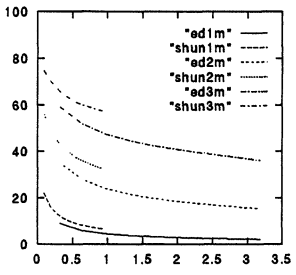


図 2: 学習量と一度しか出現しない文字列割合の関係

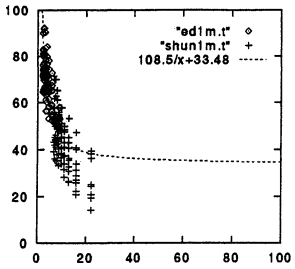


図 3: 一度しか出現しない文字列の割合と適合率の関係 (bi-gram)

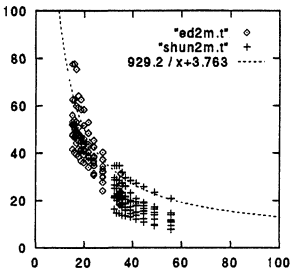


図 4: 一度しか出現しない文字列の割合と適合率の関係 (tri-gram)

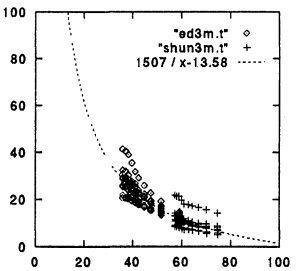


図 5: 一度しか出現しない文字列の割合と適合率の関係 (4-gram)

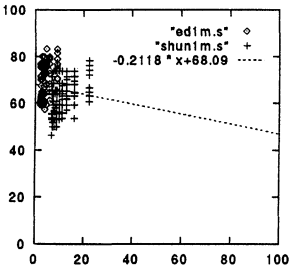


図 6: 一度しか出現しない文字列の割合と再現率の関係 (bi-gram)

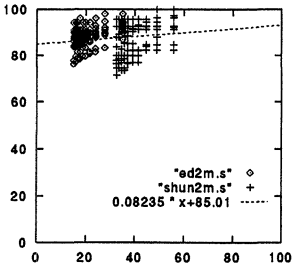


図 7: 一度しか出現しない文字列の割合と再現率の関係 (tri-gram)

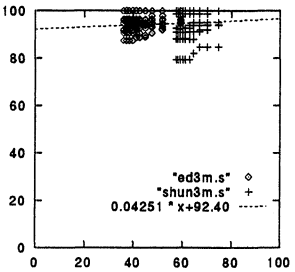


図 8: 一度しか出現しない文字列の割合と再現率の関係 (4-gram)

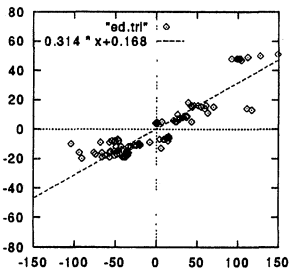


図 9: 社説での「うまく誤り指摘できた文字数」と「誤り指摘文字数」の分布 (tri-gram)

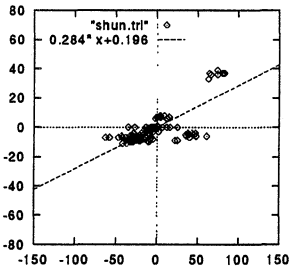


図 10: 春秋での「うまく誤り指摘できた文字数」と「誤り指摘文字数」の分布 (tri-gram)