英文曖昧検索へのHMMの適用とその評価

4P－3

太田 学　高須 淳宏　安達 淳
学術情報センター研究開発部

1 はじめに

文書画像解析と活字文字認識が実用に耐え得る精度を実現しつつあることから、文書画像をOCR認識した結果を文書データベースとして蓄える試みが行なわれるようになった。しかしOCR認識誤りの訂正コストは無視できないため、認識誤りを含むテキストをそのままデータベースに格納し、検索時にその誤りを考慮する曖昧検索手法が求められている。そこで本稿では認識前の元の文字及び2文字を状態とし、状態に遷移する際にその認識結果を出力するようなHMM(Hidden Markov Model)[1]に基づいた曖昧検索手法を提案し、英文に対する検索性能を評価する。

2 HMMを用いた英文曖昧検索手法

本稿の提案手法は、トレーニングセットに基づいて作成されたHMMを用いて1つの検索語αを複数の検索文字列βに拡張する。このとき、拡張検索文字列の妥当性を表す尺度としてHMMを元に確率$P(\beta_z|\alpha)$を求め、この確率の大きい順にランキングを行なう。その上位n個の検索文字列を用いて検索することで認識誤りによる検索洩れを改善する。

また本稿では、OCRは文字を正しく認識するか置換・欠落・挿入・結合・分解の5種類の誤りのいずれかを起こすものとする。5種類の誤りは、正しい文字(列)の長さをm、そのOCR認識結果の文字(列)の長さをnとしたとき、$\{m,n\}$の関係が、置換誤り$=\{1,1\}$、欠落誤り$=\{1,0\}$、挿入誤り$=\{0,1\}$、結合誤り$=\{2,1\}$、分解誤り$=\{1,2\}$と定めている。

2.1 作成するHMM

図1は、説明のため文字として“a”、“b”、“c”、“SP”(スペース)の4つのみを扱ったHMMの例である。作成するHMMはこのように、状態として認識前の誤りのない1文字、2文字、及び挿入誤りに対応する仮想的な文字V_iという状態をもつ。そして、文字が正しく認識されている場合及び置換・欠落・分解誤りの場合は1文字の状態に遷移し、挿入誤りの場合はV_iという状態に遷移し、結合誤りの場合はその結合する2文字の状態へと遷移する。同時に遷移先の状態を表す文字をOCRが認識した結果の文字をシンボルとして一定の確率で出力する。このとき、文字が正しく認識されている場合及び置換・挿入・結合誤りの場合は認識結果としての1文字を、欠落誤りの場合は欠落誤りに対応する仮想的な文字V_mを、分解誤りの場合はその分解した結果の2文字をシンボルとして出力する。本手法は、文字切り出しの誤りに対応するとともにbigramに基づいた文字の接続確率も考慮したものとなっている。

Applying HMM to fuzzy English-text retrieval
Manabu OHTA, Atsuhiro TAKASU, Jun ADACHI
Research & Development Department, National Center for Science Information Systems
E-mail:{ohta, takasu, adachi}@rd.nacsis.ac.jp

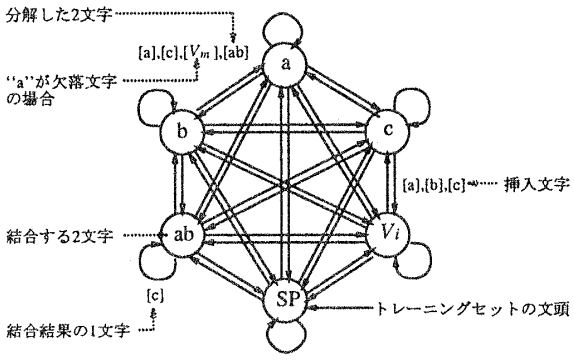

図 1: 作成するHMMの例

[正しい文字|認識結果]

[tlt][hlh][ili][sls][l][il1][sls][l][alo][l][sls][ale][m|rn][p|p][lli][elo][.|.]

図 2: HMM作成のためのテキスト

2.2 状態遷移確率とシンボル出力確率

HMMを作成するためにまず正しいテキストとその認識結果のテキストからなる英文トレーニングセットに対して、ヒューリスティクス[2]を用いて認識誤りを抽出・分類し、図2のような認識される元の文字とその認識結果の対応関係を明示したテキストを作成した。図2のテキストを入力として状態間の遷移及び遷移の際に出力するシンボルの出力頻度を数え上げることでその確率を求める。

よって状態s_iから状態s_jへの遷移回数を$C(s_i \to s_j)$として、その遷移確率$p_{s_is_j}$を次式で定義する。

$$p_{s_is_j} = \frac{C(s_i \to s_j)}{\sum_i C(s_i \to s_j)}. \tag{1}$$

一方状態s_iから状態s_jへの遷移の際にシンボルsym_kを出力する確率$q_{s_is_j}(sym_k)$は、その回数を$C(s_i \overset{sym_k}{\to} s_j)$として次式で定義する。

$$q_{s_is_j}(sym_k) = \frac{C(s_i \overset{sym_k}{\to} s_j)}{\sum_K C(s_i \overset{sym_K}{\to} s_j)} \tag{2}$$

2.3 検索語の状態系列への展開

検索語をαとすると、まず検索語となりうる状態系列S^αを以下の規則に従って全て求める。

- 最初の状態には、スペースやピリオドなどのデリミタをまとめた1つの状態をあてる。
- 2番目の状態は、検索語の1文字目の状態または、1文字目と2文字目からなる状態が存在する場合はこの2文字の状態のいずれかとなる。
- 3番目以降の状態は、前の状態の次の文字の状態、結合誤りの可能性があれば次の2文字からなる状態、挿入誤りの可能性があればV_iの状態となる。

• 挿入誤りが2文字以上続く可能性は無視する。

このような状態系列の1つを $S_x^\alpha = s^1 s^2 ... s^n$ とすると、そのとりうる確率 $P(S_x^\alpha)$ が次式で求められる。ここで最初の状態をデリミタとしているので、初期状態確率は $\pi_{s^1} = 1$ である。

$$P(S_x^\alpha) = \pi_{s^1} \prod_{i=1}^{n-1} p_{s^i s^{i+1}} = \prod_{i=1}^{n-1} p_{s^i s^{i+1}}. \quad (3)$$

このようにして状態系列とその起こる確率が全て求められると、このHMMに基づいた $P(S_x^\alpha|\alpha)$ を次式によって求めることができる。

$$P(S_x^\alpha|\alpha) = \frac{P(S_x^\alpha)}{P(\alpha)} = \frac{P(S_x^\alpha)}{\sum_x P(S_x^\alpha)}. \quad (4)$$

2.4 状態系列から出力されるシンボル系列

HMMにおける状態系列 S^α が与えられると、その状態遷移に沿ってシンボル系列 SYM^β が次式の $P(SYM^\beta|S^\alpha)$ という確率で出力される。

$$P(SYM^\beta|S^\alpha) = \prod_{i=1}^{n-1} q_{s^i s^{i+1}}(sym^{i+1}). \quad (5)$$

2.5 拡張検索文字列の得点

シンボル系列が求められると、検索語 α がある状態系列 S_x^α をとりシンボル列 SYM^β を出力する確率 $P(SYM^\beta, S_x^\alpha|\alpha)$ は、式(4)と式(5)の積で求められる。また検索語 α から SYM^β が出力される確率 $P(SYM^\beta|\alpha)$ は、そのような状態系列 S_x^α について和をとることで求められるので、結局次式が成り立つ。

$$P(SYM^\beta|\alpha) = \sum_x P(SYM^\beta|S_x^\alpha) P(S_x^\alpha|\alpha). \quad (6)$$

次に出力シンボル列の1つ SYM_y^β をそれに対応する文字列 β_z に変換する。このとき、$P(\beta_z|SYM_y^\beta) = 1$ は成り立つが、一般的には $P(SYM_y^\beta|\beta_z) = 1$ は成り立たない。よって検索文字列のランキングに必要な確率 $P(\beta_z|\alpha)$ は次式によって得られる。

$$P(\beta_z|\alpha) = \sum_y P(\beta_z|SYM_y^\beta) P(SYM_y^\beta|\alpha). \quad (7)$$

3 検索実験

Elsevierから電子形態で出版されている“Artificial Intelligence”の1995年～1998年の4年分と“Cognition”の1995年9月号～1996年6月号の論文題目と内容梗概からなるテキストデータ約440KBと約50KBを、それぞれトレーニングセット、テストセットとして英文曖昧検索を行ない、検索効率と拡張検索文字列数の関係を調べた(図3,4参照)。これらの図には比較のため、単純に認識誤りの頻度を数え上げて文字の誤る確率[1]を求め、検索語 α が β_z と認識される確率 $P(\beta_z|\alpha)$ を各文字の認識確率の積で求めた手法Iによる実験結果も示す。手法Iは文字のunigramモデルを仮定しているため接続確率は考慮されない。

[1]例えば“a”を“b”と誤る確率 $P(b|a)$ は、正しいテキストにおける“a”の出現頻度を $C(a)$、“a”を“b”と誤認識している頻度を $C(a,b)$ とすると、$P(b|a) = C(a,b)/C(a)$ で求める。

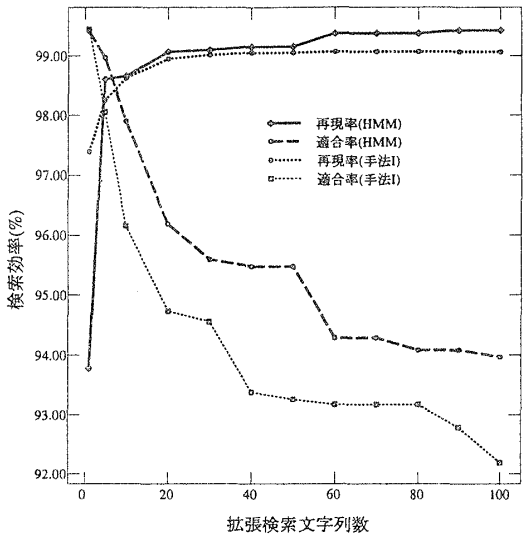

図3: トレーニングセット検索

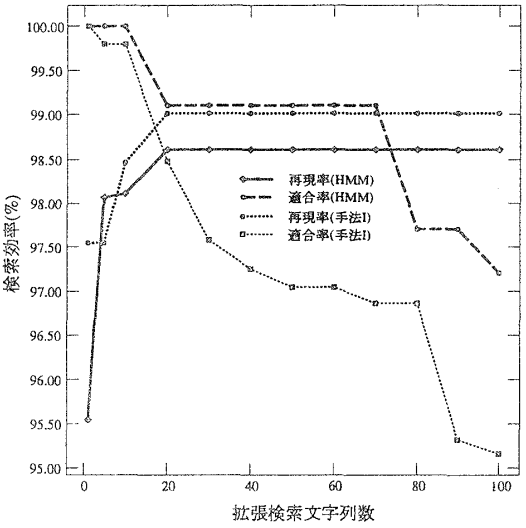

図4: テストセット検索

文字認識に用いられたOCRの認識率は98.9%、検索語拡張を行なわずに検索した場合の再現率(R)・適合率(P)はトレーニングセットでR=97.39%、P=99.43%、テストセットでR=97.54%、P=100.0%であった。

図3では提案手法は手法Iよりも全般的に良い結果が得られているが、図4では再現率が悪くなっている。

4 まとめ

提案手法は文字の接続確率を考慮しない手法よりも、トレーニングセットでは検索効率及び拡張検索文字列数の両方の点で優れているが、テストセットでは再現率の最大値の点で劣っている。これは、トレーニングセットで未観測の誤りが原因で、HMMパラメータの補間や再推定によって解決すべき問題と考えている。

参考文献

[1] Charniak, E.: *Statistical Language Learning*, The MIT Press (1993).

[2] 太田学, 高須淳宏, 安達淳: OCR認識誤りの学習方法について, 情報処理学会第57回全国大会, 1D-1(1998).