類似文字データベースとN-gramによる文字認識後処理

6Ｊ－7

平野浩次　兵藤安昭　青木恒夫　池田尚志
岐阜大学工学部

1　はじめに

市販のOCRの認識率は,カタログでは９９％以上をうたっているが,現実には印刷の品質が悪かったり,スキャナーでの読み取り条件が最適化されていないことのために,９５％程度にとどまっている.そこで,何らかの言語処理を導入した後処理が必要となる.後処理方式の多くは,N-gram統計を用いたものが多いが,この場合,文字候補の中に必ず正解文字が存在するという条件が付けられる.つまり,文字候補の中に正解文字が存在しない場合は、正解文が得られないことになる.また,正解文字がN-gramのテキストデータベースに依存する為,正しい文字を選び出さない事がある.

そこで本研究では,類似文字データベースを作成して,候補の追加を行なう方法を提案する.すなわち,OCRの候補文字ラティスにtrigramモデルを用いたコスト最小法を適用し,選ばれた最適パスの中で,ある閾値を越えた部分を誤りとして検出する.そして,その部分に類似文字データベースから新たに候補を補足して再度Ngramモデルを用いたコスト最小法を適用する.新聞記事をテストデータとして実験を行なったところ,この方法による精度の向上を確認できた.また,N-gramデータベースで使用するテキストコーパスに存在しない単語を補うため,単語辞書のN-gramデータベースを追加した実験を行った.

2　システムの概要

システムの概要を図1に示す.

OCRの結果である文字候補ラティスを入力データとして,コスト最小法により最適パスを導出する.得られたパスから誤り候補文字を抽出し,各候補に類似文字を追加,更に,コスト最小法を適用して結果を導き出す.

Postprocessing for character recognition based on N-gram model using similar character database
Koji Hirano, Yasuaki Hyodo, Tsuneo Aoki, Takashi Ikeda
Faculty of Engineering, Gifu University
Gifu-shi,501-11,Japan

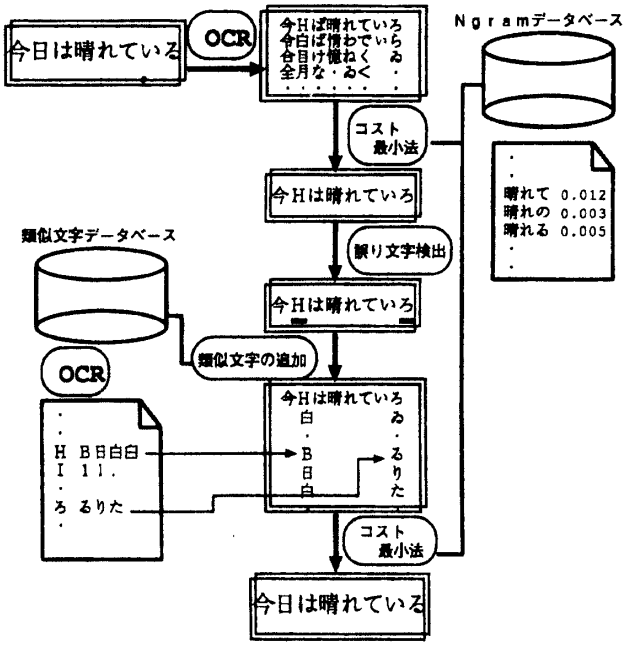

図1: システムの概要

3　N-gramデータベース及び類似文字データベース

3.1　N-gramデータベースの作成

訓練テキストとして,朝日新聞10Mbyte(約1ヵ月分)を使用し,trigram確率を求めた.ただし,bigram,unigramを使って,削除補間法[1]により線形形式の補間を行なっている.

辞書はEDRの日本語単語辞書の見出しを用いて,上記と同様にtrigram確率を求めた(表1).

表1: 学習データ数

	trigram	bigram	unigram
新聞記事	440000	130000	3000
辞書	74000	80000	4500

3.2　類似文字データベースの作成

日本語第一水準約３０００文字に対して,文字認識を行ない,正解に対する候補文字を抽出する.その候補をキーとして正解文字をレコードとするデータベースを作成する.これが類似文字データベースである.

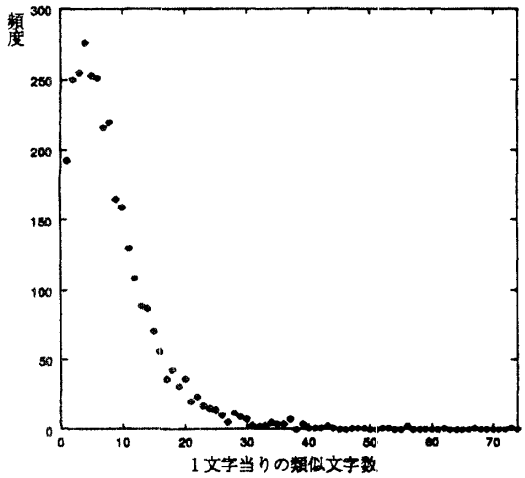

図 2: 類似文字の数の分布

類似文字数の分布を図 2に示す. 類似文字の最大数は, 7 3 文字. 最小数は 1 文字. 平均は 8 . 5 6 文字であった.

4 評価実験

4.1 実験内容

以下の実験を行なった (表 2, 表 3).

1. 類似文字データベースを使用した場合 (認識率C) と使用しない場合 (認識率B)
2. Ngram データをテキストからのみ用いる場合 (1) と, 単語辞書を追加したした場合 (2)
3. 類似文字を追加する文字を決定する閾値を変化させる (0.0001, 0.001, 0.005)
4. 学習データに用いたコーパスである朝日新聞記事 (close:表 2) と, 異なるコーパスである毎日新聞記事 (open:表 3) から無作為に抜き出したデータ約 1 0 0 0 文字づつ, をテストデータとして使用する.

なお, 後処理前の OCR の認識率 (認識率A) は, 朝日新聞は 94 % (962 文字/1015 文字), 毎日新聞は 95 % (1363 文字/1434 文字) のものを用いた. 認識率Bは,OCR が出力する候補文字のみを用いて後処理をした時の認識率である. 認識率Cは, 類似文字を追加して, 後処理をした時の最終認識率である. 再現率と適合率は誤り文字検出に対する評価を示すものである.

表 2: close テスト (朝日新聞)

	閾値	0.0001	0.001	0.005
(1)	認識率B (%)	95.6	95.6	95.6
	認識率C (%)	96.6	97.9	98.2
	適合率 (%)	60.0	33.5	21.1
	再現率 (%)	51.0	81.1	94.3
(2)	認識率B (%)	95.6	95.6	95.6
	認識率C (%)	96.5	97.7	98.2
	適合率 (%)	63.8	34.4	21.8
	再現率 (%)	43.3	79.2	94.3

表 3: open テスト (毎日新聞)

	閾値	0.0001	0.001	0.005
(1)	認識率B (%)	96.6	96.6	96.6
	認識率C (%)	96.7	96.5	96.7
	適合率 (%)	47.6	26.4	17.9
	再現率 (%)	71.8	91.5	95.8
(2)	認識率B (%)	96.6	96.6	96.6
	認識率C (%)	96.9	97.1	96.7
	適合率 (%)	46.7	29.5	18.2
	再現率 (%)	58.9	82.8	88.7

5 おわりに

従来の Ngram モデルによる文字認識後処理に, 類似文字データベースから候補を追加し, 訂正を行なう方式について述べた.

実験評価として close テストに関しては, 後処理を行うことで９４％→９８％まで認識率を向上させることができた. また、open テストに関しても、９５％→９７％まで向上した. 単語辞書データベースは,open テストでは認識率を上げるのに寄与している.

問題点としては, 訓練テキストにも辞書にも出てこない文字 (例えば漢数字の一, 十など) や語 (固有名詞など) をうまく処理できないことなどが挙げられる.

参考文献

[1] 北 研二, 中村 哲, 永田 昌明 共著 : 音声言語処理, コーパスに基づくアプローチ. 森北出版株式会社 1996.