言語情報を利用した文字認識における文字認識率と単語認識率の関係*

5N−9

大槻恭士　伊藤彰則　牧野正三　曽根敏夫
(東北大応情研)　(東北大通研)

1　まえがき

文字認識の後処理として，単語辞書や文字連接情報などの言語情報が用いられている．特に文字連接情報は，簡単で高速な処理で単語辞書と同等な効果が得られることが報告されている[1]．

本稿では，これらの言語情報を用いた文字認識における，文字認識率と単語認識率の関係を，実際に認識を行なわずに求める手法を示す．

2　文字連接情報の問題点

単語辞書 Ω から1単語 W を抽出し，残りの単語中の，W との距離が d である単語の数を $N_d(W)$ とすると，単語辞書の平均の距離 d の単語数 n_d は，

$$n_d = \frac{1}{N}\sum_{W\in\Omega} N_d(W)$$

である．切り出しが正確である場合の，文字認識率と，単語辞書を用いた単語認識率の関係は，この n_d によって求めることができる[2]．詳細は省くが，簡単にいえば n_d が大きいほど，特に n_1, n_2 が大きい(近距離単語が多い)場合に単語認識率は低くなる．

文字連接情報を満たす全ての文字列を列挙し，各単語とのハミング距離から n_d がわかれば，文字認識率と，文字連接情報を用いた単語認識率の関係を求めることができる筈である．しかし，文字連接情報の平均分岐数が B，文字列の長さが l 文字の場合，文字列の数は $O(B^l)$ と長さにつれて爆発的に増加し，n_d を求めることが困難となる．

3　文字連接情報より n_d を求める方法

そこで，以下に示すアルゴリズムで，全系列を列挙することなく n_d を求める．まず，記号の定義を示す，

(1) Ω ：認識対象（単語数 N ）
(2) $W = W_1\cdots W_L$ ：単語（長さ L 文字）
(3) $C = \{c_1,\cdots,c_M\}$ ：M 種類の文字
(4) E ：単語の終端になり得る文字の集合
(5) $A(c_i)$ ：c_i が連接可能な文字の集合
(6) $\Omega^l_{c_i}$ ：長さ l の可能な文字列の中で終端が c_i であるもの
(7) $n^l_d(W,c_i)$ ：$\Omega^l_{c_i}$ 中の，$W_1\cdots W_l$ との距離が d であるものの数

n_d は次のように定義される．

$$n_d = \frac{1}{N}\sum_{W\in\Omega}\sum_{c_i\in E} n^L_d(W,c_i)$$

そして次のような漸化式が成り立ち動的計画法で計算できる．

$$n^l_d(W,c_i) = \sum_{c\in A(c_i)} f(W,l,c,c_i,d)$$

$$f(W,l,c,c_i,d) = \begin{cases} n^{l-1}_d(W,c) & if\ W_i = c_i \\ n^{l-1}_{d-1}(W,c) & if\ W_i \neq c_i \end{cases}$$

このアルゴリズムの計算量は，時間計算量で $O(NM^2L^2)$，領域計算量で $O(LM)$ である．

4　文字認識率と単語認識率の関係

新明解国語辞典(第二版)の重要語のうち2文字以上8文字以下の4,075単語について，単語辞書を用いた場合及び文字連接情報を用いた場合の文字認識率と単語認識率の関係を求めた．文字連接情報は2,101種類の文字について，上記の4,075単語から求めた．

長さ毎の単語数，単語辞書及び文字連接情報における n_d を表1に示す．2文字の単語が圧倒的に多いのがわかる．また，単語認識率に大きな影響を及ぼす n_1 が，文字連接情報の場合でもそれほど大きくないのがわかる．特に長さが2の場合に単語辞書と文字連接情報の n_1, n_2 が殆ど同じということと，2文字の単語が多いことから，文字連接情報の有効性がうかがえる．

単語辞書を用いた場合の文字認識率と単語認識率の関係を表2に，文字連接情報を用いた場合の文字認識率

*The relationship between the character recognition score and the word recognition score on character recognition using linguistic information. by T.Otsuki, A.Ito, S.Makino(R.C.A.I.S., Tohoku Univ.) and T.Sone(R.I.E.C., Tohoku Univ.)

と単語認識率の関係を表 3 に示す．文字認識率が高い場合，文字連接情報でも高い認識率が得られ，例えば文字認識率が 90% でも，単語辞書を用いて約 99 %，文字連接情報を用いてやはり約 99 % の単語認識率が得られるのがわかる．

5 むすび

単語辞書及び文字連接情報を用いた場合の文字認識率と単語認識率の理論的関係を求め，文字連接情報が単語辞書と同程度に有効であることを示した．

参考文献

[1] 杉村, 斎藤: “文字連接情報を用いた読み取り不能文字の判定処理 - 文字認識への応用 -”, 信学論 (D), J68-D, 1, pp.64-71 (1985-1)

[2] 阿部, 秦野, 福村: “辞書を利用する文字認識系の能力の評価”, 信学論 (C), 52-c, 6, pp.305-312 (1969-6)

[3] 森健一: “パターン認識”, 電子情報通信学会 (1988)

長さ	2	3	4	5	6	7	8
単語数	3338	235	457	39	17	4	2
単語辞書の n_1	16.5	4.0	0.5	0.2	0.0	0.0	0.0
文字連接情報の n_1	19.6	40.0	40.6	39.9	5.4	17.3	15.0
単語辞書の n_2	3320.5	21.4	4.4	0.9	0.0	0.0	0.0
文字連接情報の n_2	4021.4	848.5	579.3	687.0	184.4	239.3	99.0
単語辞書の n_3	-	430.6	19.5	1.6	0.2	0.0	0.0
文字連接情報の n_3	-	26298.5	9500.0	8293.2	4266.8	1910.0	1210.5
単語辞書の n_4	-	-	209.7	4.1	1.5	0.0	1.0
文字連接情報の n_4	-	-	193633.0	98651.2	68044.9	17812.3	10473.5
単語辞書の n_5	-	-	-	31.2	2.8	0.0	0.0
文字連接情報の n_5	-	-	-	1.6×10^6	813629.7	1.7×10^5	40533.5
単語辞書の n_6	-	-	-	-	11.4	0.5	0.0
文字連接情報の n_6	-	-	-	-	1.3×10^7	1.7×10^6	3.2×10^5
単語辞書の n_7	-	-	-	-	-	2.5	0.0
文字連接情報の n_7	-	-	-	-	-	1.1×10^8	2.0×10^6
単語辞書の n_8	-	-	-	-	-	-	0.0
文字連接情報の n_8	-	-	-	-	-	-	9.8×10^8

表 1: 長さ毎の単語数，単語辞書及び文字連接情報の n_d

文字認識率	長さ 2	3	4	5	6	7	8
80	95.9	99.6	99.9	100.0	100.0	100.0	100.0
85	98.1	99.7	100.0	100.0	100.0	100.0	100.0
90	99.2	99.9	100.0	100.0	100.0	100.0	100.0
95	99.8	100.0	100.0	100.0	100.0	100.0	100.0
99	100.0	100.0	100.0	100.0	100.0	100.0	100.0

表 2: 単語辞書を用いた場合の文字認識率と単語認識率の関係 (単位 %)

文字認識率	長さ 2	3	4	5	6	7	8
80	95.0	94.9	95.2	95.2	99.3	98.0	98.3
85	97.7	97.2	97.2	97.2	99.6	98.8	99.0
90	99.0	98.6	98.6	98.6	99.8	99.4	99.5
95	99.7	99.5	99.5	99.5	99.9	99.8	99.8
99	100.0	99.9	99.9	99.9	100.0	100.0	100.0

表 3: 文字連接情報を用いた場合の文字認識率と単語認識率の関係 (単位 %)