

F-016

中途失明者用オンライン日本語入力システムにおける 重ね書き文字の自動切り出し

Segmentation of Overlapping Characters in an On-line Japanese Input System for Blind Person

長野 宏輔† 江崎 修央‡ 清田 公保†† 水野 慎士† 山本 眞司†
Kousuke Nagano Nobuo Ezaki Kimiyasu Kiyota Shinji Mizuno Shinji Yamamoto

1. はじめに

これまでに研究してきた中途失明者用オンライン日本語入力システムにおける文字の入力方法の改善を行った。従来の文字入力方法は1文字入力ごとに1文字の区切りを入力する必要があったため、筆者への負担が大きく、紙に文字を書く感覚とは大きく異なっていた。そこで、1箇所の文字枠に重ね書きで入力された文字列を自動的に分解し、認識する手法について検討した。本稿では、ペンアップ特徴量のマハラノビス距離を利用した文字の切り分けを行った後で、文字 Tri-gram や形態素解析を利用して認識精度の向上を試みた。

2. マハラノビス距離による重ね書き文字切分け

1箇所の文字記入枠に書かれた、重ね書き文字列を1文字ごとに切り出すための基本的な特徴量と、マハラノビス距離を用いた切り出し手法について述べる。

2.1. ペンアップから得る特徴量

ペンアップとは、あるストロークの書き終わりの位置から、次のストロークの書き始めの位置までの筆運びを指す。このペンアップの時間、角度、距離の3つを特徴量として利用する(図1)

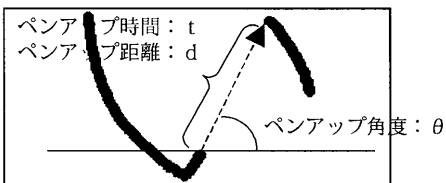


図1: ペンアップの時間、角度、距離

2.2. 認識得点から得る特徴量

1つのストロークが入力されるごとに単文字認識を行い、ストローク毎の認識得点の推移を利用する。我々の認識エンジンでは、文字が完成した時点で認識得点が一時的に高くなり、余分なストロークが入力された場合には得点が低下するという特徴がある。これを利用し、(1)式によって文字の切れ目かどうかを評価する。

$$\Delta p = ((p_i - p_{i-1}) + (p_i - p_{i+1})) / 2 \quad (1)$$

2.3. マハラノビス距離による判別分析

上述の4特徴量を用いてマハラノビス距離による文字の切れ目の判別を行う。利用する共分散行列は、サンプルデータにおける文字の切れ目でのペンアップ特徴量および文字の切れ目でない箇所でのペンアップ特徴量それぞれから求めておく。

文字入力が行われる際、それぞれの特徴量に対して切れ目と文字内のマハラノビス距離を求め、図2のようにプロットする。このとき、プロットする点が識別境界面より上部にある場合は文字内のペンアップと判断し、下にある場合は文字の切れ目であると判断する。

ここで、識別境界面は正しい切り出し箇所をもらさないことが重要であるため、切れ目となるストロークの殆どを含むように境界条件を設定した。

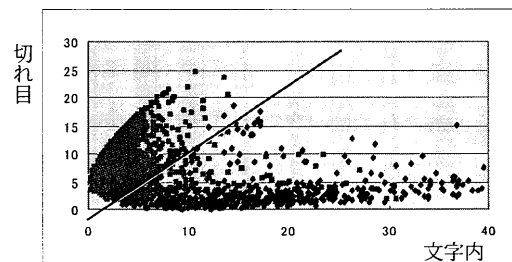
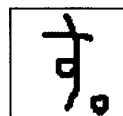


図2: マハラノビス距離の分布

2.4. 句読点の切り出し

句読点が記入されたときの特徴量が文字内のペンアップ情報と類似しているため、前節の手法では句読点は切り出しが行えない場合が多い。そこで句読点に対しては別処理を施すことにより判別を行う。図3の場合、 $S_2 > S_1$ で、かつCが句読点であった場合は、最終ストロークを句読点と判断し、切り出し箇所とする。



$S1$: [す] + [。] 部分の認識得点
 $S2$: [す] 部分の認識得点
C: [。] 部分の候補文字

図3: 句読点切り出しの例

3. 分解された文字列の結合と正解文字列の選定

3.1. 文字構成要素の結合による候補文字列作成

2章で提案した処理を、重ね書き文字列に適用すると、1文字が平均して約2個の文字構成要素に分解される。こ

† 豊橋技術科学大学

‡ 鳥羽商船工業高等専門学校

†† 熊本電波工業高等専門学校

のバラバラになった文字構成要素を、隣り合う3つの要素を限度として結合させ、候補文字列を作成する。

3.2. 文字 Tri-gram を用いた誤り訂正処理

結合された文字構成要素に文字認識および確率言語モデルの一種である文字 Tri-gram による誤り訂正処理を適用し、文字認識得点を取得する。ここで認識得点は、入力文字パターン列を $X=X_1, X_2, \dots, X_n$ 、候補文字群から生成される任意の文字列を $C=C_1, C_2, \dots, C_n$ としたときに(2)式で求められる確率値 $T(C|X)$ を指す。

$$\log T(C|X) = \log P(X|C) + \omega \log P(C) \quad (2)$$

$P(X|C)$ は、文字列 C のもとで入力パターン列 X が発生する確率で1文字認識得点を利用、 $P(C)$ は文字列 C の生成確率で文字 Tri-gram における文字列生成確率を利用する。

3.3. 単語間の文法的接続確率

3.2 節で認識されたすべての候補文字列に対して、単語辞書を利用することによって文字列を形態素に分類する。使用する単語辞書は市販の仮名漢字変換用の辞書を元に作成し、約9万2千単語が登録されている。また、サンプル文章から作成したそれぞれの品詞間の接続確率を別に用意し、分割された形態素について接続確率を(3)式により求める。

$$\log P(S) = \sum_{i=1}^n \log P(s_{i-1} s_i) \quad (3)$$

式中の $P(s_{i-1} s_i)$ は形態素列 $S = s_1, s_2, \dots, s_n$ において、形態素 s_{i-1} の品詞と形態素 s_i の品詞が接続する確率を示している。

3.4. 文中に占める構成形態素数

3.1 節の方法により結合した文字列は、全く意味を持たない単独文字が連結しているだけの文字列である場合が多い。このような文字列を排除するために形態素列を利用したペナルティ k ($k = \text{形態素数} / \text{文字数}$) を用いる。

3.5. 総当たりで得られた文字列の順位付け

候補文字列の順位付けを行う。3.2 節の Tri-gram による誤り訂正処理で得られた値を t 、3.3 節で得られた形態素同士の文法的接続の値を g 、3.4 節で得られた構成単語数の割合を示すペナルティを k とする。このとき、正解文字列候補の順位を示す得点 E の値を(4)式のように定義する。

$$E = \alpha t + \beta g - \gamma k \quad (4)$$

ここで、 α 、 β 、 γ は実験によって最適な切り出し結果が得られた値を用いた重み係数である。本稿では $\alpha = 1.7$ 、 $\beta = 0.1$ 、 $\gamma = 0.7$ とした。今回の評価実験ではこの値を利用している。このようにして、求めた正解文字列を示す得点 E の高いものから正解文字列候補として出力する。

4. 評価実験

重ね書き文字の自動切り出しの評価を行うため、目隠し暗眼者6人に手紙例文(280文節、1,522文字)を入力してもらった。これらに対し、3章で提案した各処理を施したときの結果を調査した。

4.1. 重ね書き文字切り出し位置の正解率

3章の処理を行ったときの切り出し位置の正解率を表1に示す。それぞれの結果は、文字 Tri-gram による誤り訂正処理の認識得点で評価したもの、単語間の文法的接続確率を加えて評価したもの、さらに文中に占める形態素数を加えて評価したものである。

表1 重ね書き文字の切り出し正解率 (3章の処理)

	一位	累積二位	累積三位
Tri-gram 処理のみ	87.9%	94.9%	96.0%
+文法接続確率	92.0%	95.0%	96.8%
+形態素数	92.6%	95.9%	96.9%

表1から、文字構成要素をすべてのパターンで結合させた後、文字認識、文法的接続確率、形態素数を利用することによって、より正しい位置で切り出された文字列を選択できることが分かる。

4.2. 文字認識率の比較

従来手法と提案手法との文字認識率の比較を行った。従来手法は、重ね書き文字データに正解切り出し位置を与えて文字認識を行い、Tri-gram を用いた訂正処理の認識得点により順位付けを行った。提案手法は、Tri-gram による誤り訂正処理の認識得点 ((4)式、 αt の項まで)、文法的接続確率を加えた得点 (同 βg の項まで)、文字列に占める形態素数を加えた得点 (同 γk の項まで) でそれぞれ順位付けし、文字認識率を求めた。

表2 重ね書き文字の自動切り出し認識結果

	一位	累積二位	累積三位
従来入力法	93.1%	94.7%	95.2%
Tri-gram 処理のみ	83.7%	88.0%	90.0%
+文法接続確率	83.9%	88.3%	90.4%
+形態素数	85.4%	89.5%	91.1%

表2の結果から、3章で述べた3つの特徴が、すべて認識精度の向上に貢献していることが分かる。そして、提案手法による最終的な文字認識率と従来入力法による認識率を比較してみると、1位認識率で約7%、累積2位、3位認識率で約5%の精度低下が見られることが分かった。

5. まとめ

ペンアップ情報および、文字認識得点の推移を使用した判別分析によって重ね書き文字列を分解した。そして、分解した文字列を再結合し、日本語らしさを基準とした順位付けによって正解文字列候補を求めた。この処理によって、重ね書きによる文字入力が可能となり、その文字認識率は従来文字入力法と比較して約7%の低下となった。

今後の課題は、文法規則による接続確率を品詞 n-gram に置き換えることで認識精度を向上させること、また、処理の高速化と認識失敗したときの訂正方法の検討などが挙げられる。