2C−1

オートマトン型単語照合の姓名文字列への適用

丸川勝美　古賀昌史　嶋好博　藤澤浩道
(株)日立製作所　中央研究所

1．まえがき

オフィスのオートメーション化に伴い、大量の情報を入力することが必要であり、また、氾濫する情報を整理する基準の一つとして「だれが」（姓名）が重要となっている。そのため、姓名を認識し自動的かつ効率よく整理することが要求されており、文字認識や音声認識等が注目されている。

文字列入力の際に生じる曖昧さを解消し精度向上を図るため、辞書中の単語を利用し照合する単語照合方法がある[1]。従来、認識候補を組み合わせた文字列と辞書中の単語を照合させる方法がある。この方法は認識候補を組み合わせるため、正解の文字が候補中に存在しないと正解姓名を得られない。また、認識候補の組み合わせの爆発を起こさないように、候補数を限定するといった不具合がある。そのため、これらを解決したオートマトン型単語照合が提案されている[2][3]。この方法は上記の課題を解決し、膠着した文字列から任意の位置の単語を実時間で抽出することが可能である。

本報告では、姓名文字列を高精度に入力させる際の特徴、および、これをオートマトン型単語照合へ適用した時の結果とその考察について述べる。

2．姓名文字列照合の特徴

ユーザが自由に姓名を記入でき、その照合処理を高精度で実時間内に行なう必要がある。ここで、ユーザが姓名を記入する際の方法として、

①分ち書き姓名
②べた書き姓名

の2つがある。①は姓と名の間をスペース等で区切って記入する方法である。また、②のは姓と名を連続して記入する方法である。

これらの記入に依存しない方式にするには、任意の桁位置に記入された姓そして名を実時間で処理する必要がある。例えば、図1に示すように、認識装置は記入文字列の各文字に対し複数個の候補文字（N個）を出力するため、4×Nの候補文字ラティスから正解文字列を推定することになる。また、照合を行なう単語辞書が大規模であるため、いかに高速に処理を行なうかが特徴となる。

また、認識装置は必ずしも正解の文字を候補中に出力しない。そのため、正解の文字が存在しなくても正解の単語を抽出する必要がある。

さらに、姓名文字列は「大井」「大田」や「章子」「祥子」のように2文字の単語であるが1文字しか異ならない単語や1文字の単語が存在する。そして、姓名をべた書きで記入した場合、姓と名の切れ目が明確でない姓名文字列が存在する。

本姓名照合は、図1に示すように生成された候補文字ラティスに対し、桁ずらしによるオートマトン型単語照合を用い、実時間で姓・名候補を抽出する。

3．オートマト型単語照合

オートマトン型単語照合は候補文字ラティスから有限オートマトンを生成し、辞書単語

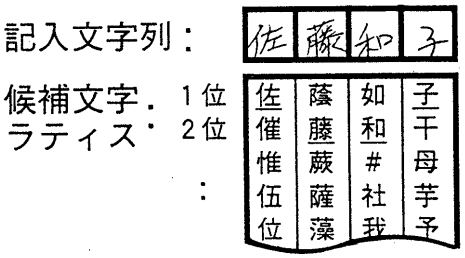

図1　候補文字ラティス

An Application of an Automaton-type Word Matching to a Name String
Katsumi MARUKAWA, Masashi KOGA, Yoshihiro SHIMA, Hiromichi FUJISAWA
Central Research Laboratory, Hitachi,Ltd.

を構成している文字でオートマトンの状態間を遷移し、その確信度を求めるものである。

単語照合で用いるオートマトンを、図２を用いて説明する。候補文字ラティスから生成したオートマトンは、状態の遷移を１文字先に限定したもので、状態数は記入文字数に１を加えたものである。また、各状態間の枝の本数は、各文字の桁位置における候補文字数に１を加えたものである。そして、それぞれの枝には候補文字の圧縮変換コードと候補文字の順位に相当したペナルティが与えられる。状態の遷移は、照合開始桁から順に辞書単語を構成している文字の圧縮変換コードを一つづつ入力し、そのコードに対応する枝を通り、ペナルティを積算しながら次の状態へと移る。圧縮変換コードに相当した枝が無い場合、「その他」の枝を通りペナルティとしてＰを積算する。そして、ペナルティを積算したものが単語距離となる。例えば、辞書単語「佐藤」が状態０から入力されると単語距離は１(＝０＋１)となる。

桁ずらしによるオートマトン型単語照合はべた書き文字列の任意の桁位置にある単語を抽出するためのものである。その原理は、オートマトンに入力する辞書単語との照合を開始する状態を照合開始アドレスにより制御し、任意の桁位置において単語照合が行なえるようにしたものである。また、照合開始アドレスが指している状態を照合開始桁とし、候補文字を基に探索したすべての辞書単語に対して単語照合を行なう。照合が終了すると、照合開始アドレスを次の状態へと更新する。そして、同様な操作により、新たな照合開始桁で単語照合を行う。このように照合開始アドレスを１状態づつ更新することで一桁づつ照合開始桁をずらしながら単語照合を行い、任意の桁位置に記入された単語を抽出することが可能となる。

４．実験結果

実験はワークステーション2050/32(cpu:MC68020､20MHz)で行い、Ｃ言語で記述した。

処理時間を調べるため、姓と名が別々に記入されたと仮定し、姓：１３９１サンプルそして名：１０９１サンプルに対する大量実験を行い、処理時間および入力文字列の曖昧さの解消の度合いを実験で求め、良好な照合結果を得た。照合が大きな効果を出さない対象は、１文字単語や正解の文字が候補中に存在しい場合であった。

５．あとがき

ユーザが自由に記入した姓名文字列の認識精度を向上させるため、姓名文字列を認識する際の特徴およびオートマトン型単語照合の適用について報告した。今後はフリガナによる高精度化の方法について検討したい。

参考文献

[1] 藤澤、中野、安田：漢字認識における文脈情報利用の一方式、昭５２信学総全大、S10-3

[2] 丸川、古賀、嶋、藤澤：オートマトン型を用いた曖昧文字列からの単語抽出方法、平４春季信学全大、D-110

[3] K.Marukawa, et al. : A High Speed Word Matching Algorithm for Chinese Character Recognition, MVA'90 IAPR Workshop, pp.445-449, 1990

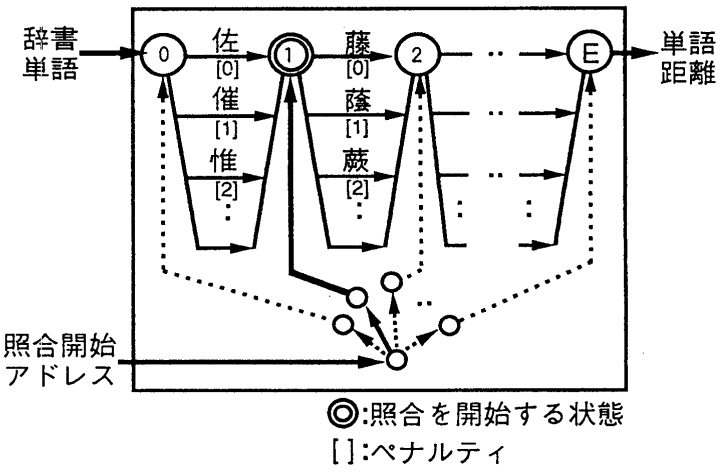

図２ オートマトン型単語照合の構成図