4L−1　オートマトン型単語照合の高速化手法

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

知的なヒューマン・インタフェースの実現のため、日本語入力方法として文字認識や音声認識等による文字列の入力が注目されている[1],[2],[3]。文字列の入力では入力作業の効率化・自動化が求められており、文字列を高速に入力することが強く要求されている。

文字列入力の際に生じる不確実性を解消するため、単語辞書を利用し、単語を照合する単語照合方法がある。この方法は辞書の単語と候補文字ラティスを比較することで、意味を持った候補文字列を抽出し、正解文字列を推定するものである。この単語照合方法として、既に、有限オートマトン(FSA)を利用したオートマトン型単語照合手法[4]を提案している。入力文字列が単語に分かれていないべた書き文字列や連続した音声の入力において、単語の区切りがないために処理量が膨大となり、単語照合の高速化が要望されている.

本報告では入力文字列一文字に対し一回の遷移を行い、また、状態遷移テーブルの表現方法を工夫することで、従来手法を高速に処理する改良型の手法を新に提案する。

2．高速オートマトン型単語照合手法

本手法は漢字を対象とし、２バイトで表現される文字コードを圧縮変換した変換コードを利用する。そして、このコードにより、入力文字列一文字に対し一回の遷移を行うことで状態遷移時間を削減する。また、この変換コードを用いた状態遷移テーブルの表現形式を工夫している。これにより、状態遷移テーブルの初期化処理を高速に行う。

図１に示すオートマトンに対し、状態１に一文字入力すると状態１から状態mに遷移し、次の文字の入力で状態nに遷移することを示している。状態１から状態mに遷移するコードはS_1、S_2、S_3、otherのいずれかであり、遷移する際のコストは各々C_1、C_2、C_3、C_0である。また、状態mから状態nに遷移するコードはT_1、T_2、otherの一つであり、コストは各々C_1'、C_2'、C_0'である。例えば、状態１でコードS_1が入力すると、遷移先はmとなり、コストC_1が積算される。そして、状態mでコードT_2が入力すると、遷移先はnとなり、コストC_2'が積算される。

高速オートマトン型単語照合手法の全体フローは、状態遷移テーブルの初期化ルーチン、状態遷移テーブルおよびコストテーブルの生成ルーチン、辞書単語との照合を行うコスト計算ルーチンから構成される。このうち、コスト計算ルーチンでは、単語辞書中の単語をオートマトンに順次入力し照合を行なう。

3．実験結果

実験はワークステーション上で行なった。使用したCPUは68020(20MHz)であり、Ｃ言語で記述した。

状態遷移テーブルの初期化、一文字当りの状態遷移テーブルおよびコストテーブルの生成時間、また、一文字遷移当たりのコスト計算時間を実測した。その結果、従来のオートマトン型単語照合手法[4]に比べ4.9倍の高速化を実現している。さらに、入力文字列の不確実性の解消の度合いを実験で求め、良好な照合結果を得た。

4．あとがき

単語に分かれていないべた書き文字列である日本語を実時間かつ高精度にで入力することを目的とし、高速オートマトン型単語照合手法を提案した。この手法は、従来の手法に比べ4.9倍の高速な処理が行えることを実験により確認した。

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]畑岡，天野，市川：概念ネットワークを用いた大語彙単語音声認識、信学論(D)，J72-D-Ⅱ，8，pp.1270-1275(昭64-8)

[3]E.M.Riseman,A.R.Hanson,A contextual Postprocssing System for Error Correction Using Binary n-Grams, IEEE Trans. on COMPUTER,Vol.C-23,No.5,MAY,pp.480-493,1974.

[4]丸川，古賀，嶋，藤澤：文字列入力のためのオートマトン型単語照合手法の一考察，平2情処春季全大，D-80

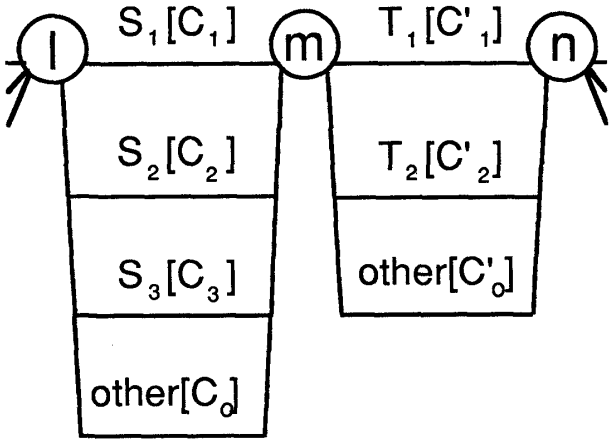

図１　高速単語照合のためのオートマトン

A High Speed Algorithm for Automaton-type Word Matching

Katsumi MARUKAWA, Masashi KOGA, Yoshihiro SHIMA, Hiromichi FUJISAWA

Central Research Laboratory, Hitachi,Ltd.