自然言語処理 114－15
(1996. 7. 19)

DFAによる形態素解析の高速化

森 信介
京都大学工学研究科
〒606-01 京都市左京区吉田本町
mori@kuee.kyoto-u.ac.jp

あらまし
入力文を単語に分割し品詞を付加する形態素解析は、日本語処理における基本的な処理である。日本語には単語間に明確な区切り記号がないので、この処理は入力文の全ての部分文字列に対する辞書検索を含む。本論文では、辞書を決定性オートマトンに変換し、辞書検索を高速に実現する方法を提案する。この方法は、AC法(失敗関数を持つトライ)に基づく方法と比較して、計算時間が少ないという利点と、大きい記憶域を必要とするという欠点がある。これらの方法を実装し実験を行なった結果、決定性オートマトンによる方法はAC法に基づく方法に対して、必要な記憶域は16.1倍であり、辞書検索の速度は11.7倍であった。
キーワード 形態素解析 辞書検索 高速化 決定性オートマトン AC法

High Speed Morphological Analysis using DFA

Shinsuke Mori
Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-01 Japan
mori@kuee.kyoto-u.ac.jp

Abstract
Morphological analysis, which segments the input sentence into words and attaches parts of speech to them, is the most fundamental process of Japanese language processing. This process contains dictionary look-up of all substrings of input sentence. In this paper, we propose a method to convert the dictionary into a deterministic finite automaton and realize high-speed dictionary look-up. An advantage of our method is that it enables faster dictionary look-up and a disadvantage is that required memory space is larger than AC method-based dictionary look-up. The experimental results tells that our method requires 16.1 times as large memory space as AC method and is 11.7 times as fast as AC method in dictionary look-up.
Key Words Morphological analysis, Dictionary lookup, Speedup, DFA, AC method

1 はじめに

入力文を単語に分割し品詞を付加する形態素解析は、日本語処理における基本的な処理である。このため、今日までにさまざまな研究がなされ、標準的な解析方法が確立するとともに、実用に耐え得る解析精度が達成されている。その結果、情報検索や自然言語の機械学習の研究の多くに形態素解析が使われている。形態素解析の速度向上は、より高速な情報検索の実現や、より大きいコーパスを用いた機械学習や知識獲得の研究に有効であると考えられる。

標準的な形態素解析の処理では、入力文の各位置から始まる任意の長さの部分文字列が辞書に含まれるか否かを調べ、それぞれの部分文字列の間の接続コストや接続確率を調べ、すべての可能な解の中で確率やコストの基準で最適な解を入力文の解析結果とする。これらの処理のうち本論文で問題とするのは、辞書検索である。

辞書検索の標準的方法は、トライを用いる方法 [1] である。この方法は、各位置から始まる部分文字列はすべてその中の最長の文字列の接頭辞であるという性質を利用しており、入力文の各位置において 1 回の辞書検索でその位置から始まる全ての形態素を出力する。この方法の他の利点として、辞書を記憶しておくために必要な記憶域が少なくてすむことが挙げられる。青江 [2] は記憶域をさらに減らすために、二つのトライを用いる方法を提案している。颯々野他 [3] はこれを形態素解析に用いた結果を報告している。

丸山 [4] は、トライを用いる方法では複数回読まれる文字があるという無駄を指摘し、Aho と Corasick [5] が提案した文字列マッチングの方法を形態素解析の辞書検索に応用し高速化を実現している。これはトライの各ノードに失敗関数を持たせることで実現される。この方法の記憶域はトライと同程度であり、速度は理論的にトライを上回ることが保証されているので、トライを用いる方法より優れている。以下では、この方法を AC 法と呼ぶ。

本論文では、トライを決定性オートマトンに変換し、AC 法よりも高速な辞書検索を実現する方法を提案する。この方法 (以下 DFA 法) は、AC 法と比較して、計算時間が少ないという利点と、大きい記憶域を必要とするという欠点がある。従って、どちらの方法が良いかは計算機資源などの制約に依存する。本論文では、これらの方法を計算時間と記憶域という点で比較する。

速度の点では、辞書検索の部分を切り離して考える限り、DFA 法よりは速くならないことが以下のようにして示される。文字列マッチングの問題では、少なくとも入力文字列の長さの亜線形の時間[1]が必要であることが証明されている [6]。形態素解析においては、1 文字からなる形態素が辞書に含まれていると仮定して良いので、全ての文字を最低でも 1 回以上読む必要がある。本論文で述べる DFA 法は全ての文字を 1 回しか読まないので。形態素解析の辞書検索において DFA 法は理論的に最も高速なアルゴリズムである。

以下の節では、まず AC 法に基づく方法について述べる。次に、DFA 法に基づく方法について説明する。さらに、それぞれの長所と短所を実験の結果とともに論じる。最後に、本研究の結論と今後の課題を述べる。

2 AC 法に基づく方法

この節では、高速な形態素解析のための辞書構造として丸山 [4] が提案した方法 (AC 法) における辞書の構成と検索について述べる。

2.1 辞書構造の概要と検索方法

AC 法による辞書は、次の 6 項組からなる出力付きのオートマトンである。

1. Q は状態の有限集合
2. Σ は入力アルファベットの有限集合
3. g は $Q \times \Sigma$ から $Q \cup \{\mathrm{fail}\}$ への写像 (前方関数)
4. h は Q から Q への写像 (失敗関数)
5. q_0 は初期状態
6. F は Q から形態素情報への写像 (出力関数)

このオートマトンの状態は登録語の表記から得られるトライのノードに対応している。初期状態はルートノードに対応する。前方関数 g は、各ノードから子ノードへの写像であり、1 文字を読み込んだ後の遷移先を与える。失敗関数 h は、前方関数 g が定義されていない (g=fail) 場合の遷移先を与える。辞書検索のときは、まず 1 文字読み込み前方関数 g が定義されている状態に到達するまで失敗関数 h で 0 回以上遷移したあと、前方関数 g のその文字に対する値へ遷移する。その後、出力関数 F で与えられる形態素情報を出力する。出力される形態素は、初期状態からの経路として与えられる文字列のすべての接尾辞を表記としてもつ形態素の情報である。これは、確率的形態素解析では、文字数と品詞番号と頻度の組である。

例はば、図 1のようなトライから図 2のような AC 法の辞書が構成される。AC 法の辞書では、トライにノード間の親子関係を表わす実線が前方関数 g に対応しす

[1] 計算時間が長さ n に比例し、係数が 1 未満であること。

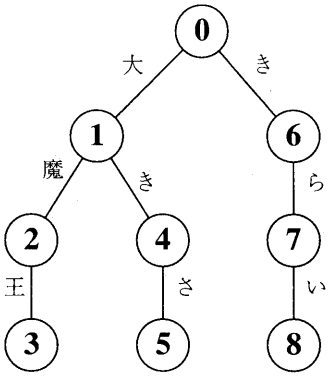
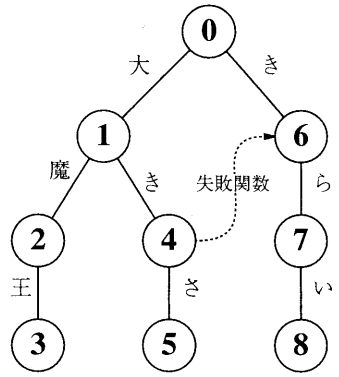


図 1: トライを用いた辞書の構造



ノード 4 以外の失敗関数は省いてある。

図 2: AC 法の辞書の構造

る。これに加えて、点線で与えられている失敗関数 h が各ノードに与えられる。状態 3 の出力関数は、「大魔王」と「魔王」と「王」がそれぞれ名詞であるとすると、$F = \{(3, \text{名詞}), (2, \text{名詞}), (1, \text{名詞})\}$ である。形態素解析のときには文脈が与えられているので、長さから表記が復元できることに注意しなければならない。この例の辞書で文字列「大きらい」に対して辞書検索を行なうと、初期状態から次ような状態遷移と出力を行なう。

$$
\begin{aligned}
&g(q_0, \text{大}) = q_1, \quad && F(q_1) = \{(1, \text{接頭辞})\},\\
&g(q_1, \text{き}) = q_4, \quad && F(q_4) = \{(2, \text{形容詞})\},\\
&g(q_4, \text{ら}) = \text{fail}, \quad && h(q_4) = q_6,\\
&g(q_6, \text{ら}) = q_7, \quad && F(q_7) = \{(1, \text{接尾語})\},\\
&g(q_7, \text{い}) = q_8, \quad && F(q_8) = \{(3, \text{形容動詞})\}.
\end{aligned}
$$

トライを用いる方法と異なり、出力される形態素は最後に読み込んだ文字の位置で終る形態素の集合である。

実装に際しては、前方関数 g は文字コードと遷移先の対の列で実現した。失敗関数 h の値は、この列の最後にアルファベットにはない文字を表わすコード (番兵) と遷移先の対を付加することで実現した。従って、遷移先はこの列をシーケンシャルに調べていき、入力文字にマッチする箇所が見つかるか、または見つからずに番兵に到達して、それに対応する遷移先を読みとることで決定される。同じことを、ハッシュや二分探索などで実現することも可能であるが、多くのノードでは数個程度の文字に対してのみ前方関数が定義されているので、逆に処理時間がかかると考えられる。出力関数 F は、形態素情報の列の先頭を指すポインターの配列で実現した。

2.2 辞書の構築

まず、与えられた形態素の表記の集合に対してトライ [7] を作成する。これによって、前述の 6 項の関数 h, F 以外の項が決定される。出力関数 F の値は、そのノードに対応する文字列の接尾辞のそれぞれに対して表記が一致する形態素の情報の集合である。失敗関数 h の値は、以下のようにして計算される。ただし、$\text{depth}(q)$ はノード q の深さを返す関数とする。

$h(q_0) := q_0$
foreach q $(\text{depth}(q) = 1)$
    $h(q) := q_0$
foreach d $(1, 2, \cdots)$
    foreach q $(\text{depth}(q) = d)$
        foreach p $(g(q, \exists c) = p)$
            $h(p) := f(h(q), c)$

関数 $f(q, c)$ の定義は以下の通り。

$$
f(q,c) = \begin{cases} h(q) & (g(q,c) \neq \text{fail}) \\ f(h(q),c) & (g(q,c) = \text{fail} \ \wedge \ q \neq q_0) \\ q_0 & \text{その他の場合} \end{cases}
$$

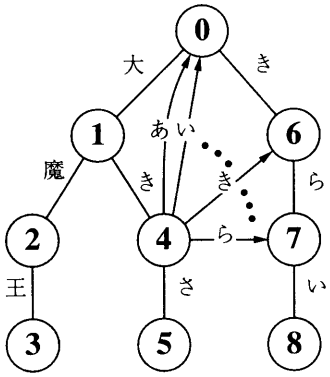
3 DFA に基づく方法

この節では、形態素解析のためのトライ辞書を決定性オートマトンに変換する方法とその検索について述べる。

3.1 辞書構造の概要と検索方法

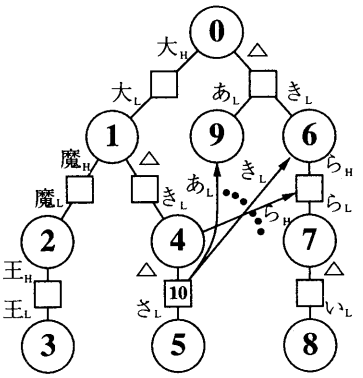
DFA 法による辞書は、次の 5 項組からなる出力付きのオートマトンである。

1. Q は状態の有限集合
2. Σ は入力アルファベットの有限集合
3. δ は $Q \times \Sigma$ から Q への写像 (遷移関数)
4. q_0 は初期状態
5. F は Q から形態素情報への写像 (出力関数)



ノード4以外に付加される矢印は省いていある。

図 3: DFA 法の辞書の構造



ノード4と10以外に付加される矢印は省いてある。

図 4: アルファベットを分割する DFA 法の辞書の構造

遷移関数 δ 以外の4項は、AC法とまったく同じである。遷移関数 δ は、各状態で1文字読み込んだ後の遷移先を与える。AC法との違いは、状態とアルファベットの全ての組み合わせに対して遷移関数 δ が定義されていることである。AC法における、失敗関数 h による0回以上の遷移と前方関数 g による1回の遷移が、これによる1回の遷移に対応する。出力関数 F の定義はAC法の場合と同じである。

例えば、図1のようなトライに、子ノードへのリンクのラベル以外の全てのアルファベットに対して遷移先を矢印のように決定することで、図3のオートマトンが得られる。この例の辞書で文字列「大きらい」に対して辞書検索を行なうと、初期状態から次ような状態遷移と出力を行なう。

$$\delta(q_0, \text{大}) = q_1, \quad F(1) = \{(1, \text{接頭辞})\},$$
$$\delta(q_1, \text{き}) = q_4, \quad F(4) = \{(2, \text{形容詞})\},$$
$$\delta(q_4, \text{ら}) = q_7, \quad F(7) = \{(1, \text{接尾語})\},$$
$$\delta(q_7, \text{い}) = q_8, \quad F(8) = \{(3, \text{形容動詞})\}.$$

AC法と同様、出力される形態素は最後に読み込んだ文字の位置で終る形態素の集合である。

この例からも分かるように、DFA法における状態遷移はどの入力文字に対しても必ず1回であり、出力関数に関係する部分はAC法と全く同じなので、AC法よりも高速に辞書検索が行なえる。この点に加えて、どの入力文字に対しても遷移関数が定義されていることから、これをアルファベットを添字とする配列を用いて実現できるので、遷移先の参照がAC法と比べて高速に行なわれる。

3.2 辞書の構築

AC法における辞書の構築と同様に、まず与えられた形態素の表記の集合に対してトライを作成する。これによって、前述の5項のうち、関数 δ, F 以外が決定される。出力関数 F の作成はAC法の場合と全く同じである。関数 δ の値は、以下のようにして計算される。ただし、string(q) は状態 q に対応する文字列を返す関数(1対1写像)であり、findLS(s) は文字列 s のトライに対応する状態をもつ最長の接尾辞を返す関数とする。

```
foreach q (q ∈ Q)
    foreach c (c ∈ Σ)
        s := findLS(string(q) · c)
        δ(q, c) := string^{-1}(s)
```

日本語のように文字数が多い言語にこの方法をこのまま適用すると、各ノードの遷移関数の記述長が非常に長くなり、結果として膨大な記憶域が必要となる恐れがある。実際、次節で述べる実験で用いたコーパスに出現する92,648個の異なり形態素をトライに変換すると、そのノード数は156,917であった。アルファベットの大きさを10,000とし、ノードの識別に4[byte]必要であるとすると、全てのノードに対する遷移関数の記述に必要な記憶域は $156,917 \times 10,000 \times 4[\text{byte}] \approx 5.85[\text{Gbyte}]$ となる。この大きさは現在の計算機環境を考えると現実的とは言い難い。この問題を回避するために文献[8]には、アルファベットを別のアルファベットの組合せとして表わす方法が紹介されている。例えば、日本

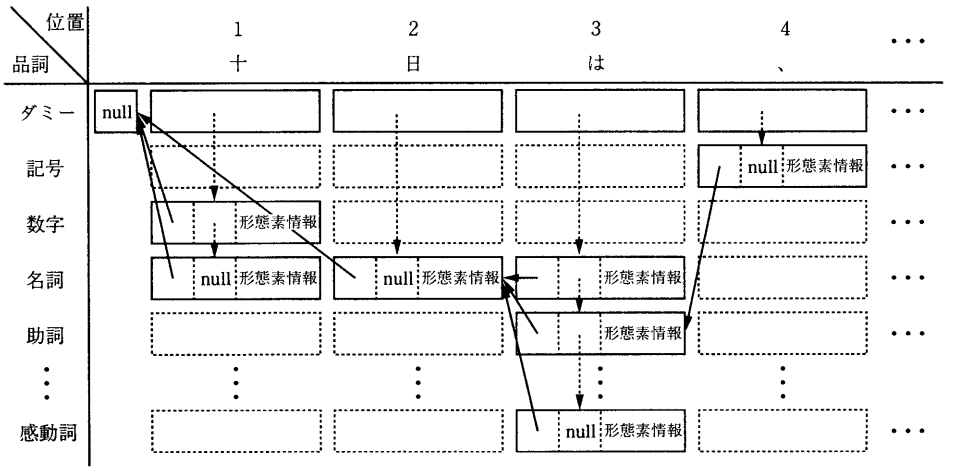


図 5: 最尤解の探索に用いる表

語の文字コードを上位バイトと下位バイトに分割して100 個のアルファベットで表わすと、状態数は 2 倍になるが、各ノードの遷移関数の記述に必要な記憶域が1/100 になるので、遷移関数全体の記述に必要な記憶域は $156,917 \times 2 \times 100 \times 4$[byte] ≈ 119.7[Mbyte] となり、現在の計算機環境で実現可能な大きさとなる。これにより、図 3のオートマトンは図 4のようになる。この変更は、必要な記憶域を減少させる効果がある反面、1 文字の入力に対して 2 回の遷移が必要になるため、辞書検索の速度が低下する点に注意しなければならない。次の節で述べる実験では、この改良を採用している。

4 実験結果とその評価

この節では、2節で述べた AC 法と 3節で述べた DFA 法を実装し、辞書検索や形態素解析の時間的・空間的性能について実験した結果を提示し、その評価を与える。

4.1 実験環境

実験には品詞 bigram を用いた確率的形態素解析システムと EDR コーパス [9] を用いた。これを学習コーパス (103,901 文、 2,552,332 形態素) とテストコーパス (103,901 文、 2,552,715 形態素) に分割した。空間的性能は、学習コーパスから得られる辞書の記述に必要な記憶域で評価し、時間的性能はテストコーパスの辞書検索や形態素解析に要する時間で評価した。ただし、頻度 2 以上の形態素のみ辞書に記述した。実験に用いた計算機は、SPARC Station 20 (70MHz) である。辞書の構築に使用したプログラミング言語は Perl5 であり、辞書検索や形態素解析に使用したプログラミング言語は C++ である。

形態素解析の解探索の方法は、ダイナミックプログラミングを用いる標準的な方法である。これに用いる表の横方向は入力文中の位置に対応し、縦方向は品詞に対応する (図 5参照)。表の各要素はトレリスのノードに対応しており、表中の位置で示された文字と品詞で終る形態素の中で最大の確率を与える形態素と、探索のために便宜的に与えられたポインターを 2 つ持っている。1 つは、そのノードへの最尤の経路における一つ前のノードを指しており、表を埋めた後の最尤の経路の探索に用いられる。もう 1 つは、その位置で終る他の品詞のノードを指しており、このポインターをたどることで、この位置での接続確率のチェック対象を存在する品詞だけに限定することが可能になる。図 5の例の場合、入力文の 1 文字目で終る形態素の中で最尤経路の探索の対象となるのは、表の 1 文字目の欄を上から順にポインターをたどることで、「十 / 数字」と「十 / 名詞」だけであると分かる。

テストコーパスの解析には、未知語 (学習コーパスに現れなかった形態素) の処理が必要である。一般的に、日本語の形態素解析の場合は、文字種などの情報を用いたパターンマッチングのような方法を使う。本研究で用いた未知語モデルは、1 つの記号を品詞が「記号」である形態素とみなし、任意長の数字の連接を品詞が「数字」である形態素とみなし、アルファベット、平仮名、片仮名、漢字のそれぞれの任意長の連接を品詞が「名詞」である形態素とみなす。この未知語モデルの実装は、文字種に対応する状態をもつオートマトンと、同じ文字種が何文字連続したか

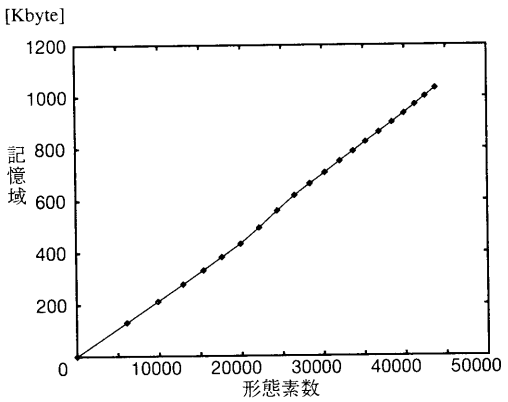


図 6: AC 法における形態素数と記憶域の関係

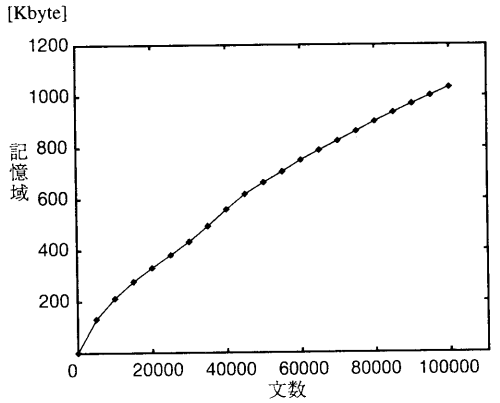


図 7: AC 法における文数と記憶域の関係

を記憶するカウンタから構成されている。この未知語モデルは、記号を除く同一文字種からなる文字列に対して、全ての部分文字列を形態素とみなすので、部分的に既知語を含む同一文字種からなる文字列 (特に片仮名列や漢字列) の解析に有効であると考えられるが、膨大な数の未知語が接続を調べる対象となるため、大幅な速度低下の原因になる。解析速度という観点からは、最長の同一文字種からなる最長の文字列だけを対象にするなどの方法が有効であるが、形態素解析の未知語処理のためにどのようなモデルが最適であるかは、解決すべき課題である。

4.2 実験結果とその評価

学習コーパスから計算されるデータの中で、 AC 法と DFA 法で異なるのは、辞書構造の記述部分だけである。その他のデータ、すなわち AC 法と DFA 法で共通のデータとその記憶域は以下の通りであった。

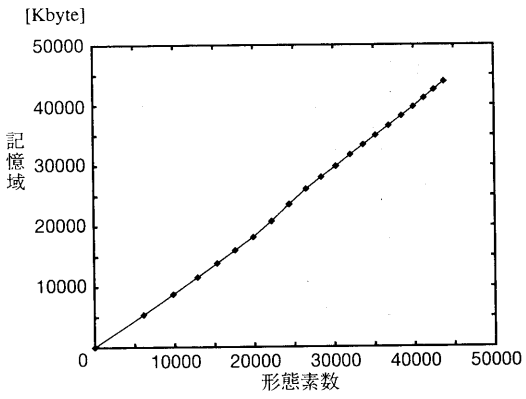


図 8: DFA 法における形態素数と記憶域の関係

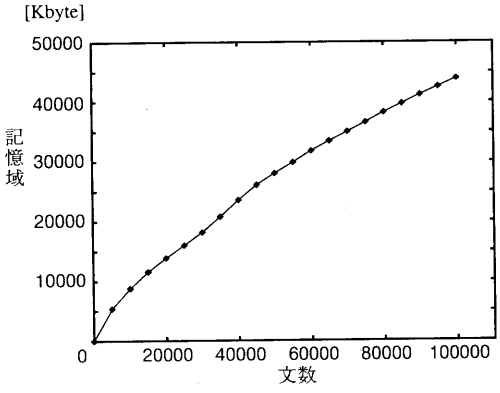


図 9: DFA 法における文数と記憶域の関係

1. 品詞番号と品詞名の対応表 (162[byte])
2. 遷移確率行列 (1,096[byte])
3. 辞書の出力情報 (1.7[Mbyte])

速度の測定は、これらを主記憶に載せて行なった。

AC 法を用いて、学習コーパスから辞書や遷移確率などを計算するの要した時間は 30 分 31 秒であった。辞書の構造の記述に要した記憶域は 1.1[Mbyte] であった。さらに、学習コーパスの文数を変化させ、構造の記述に要する記憶域の変化を調べた。図 6は横軸に (異なり) 形態素数をとった場合のグラフであり、図 7は横軸に文数をとった場合のグラフである。図 6を見ると、辞書の記憶域は形態素数にほぼ比例することが分かる。図 7から、文数が増加するにつれて辞書の記憶域は増加するが、その増加率が減少していくことが分かる。これは、文数に対して (異なり) 形態素数の増加率が減少していくことの帰結であると推測される。

DFA 法を用いた場合、辞書や遷移確率などの計算時間は 2 時間 25 分 1 秒であった。辞書の構造の記述に要した記憶域は 43.9[Mbyte] であった。 AC 法の場合と同様に、学習コーパスの文数を変化させ構造の記述に要する記憶域の変化を調べた。図 8は横軸に (異なり) 形態素数をとった場合のグラフであり、図 9は横軸に文数をとった場合のグラフである。これらの図から、文数や形態素数と記憶域の関係について AC 法の場合と同じことが言える。

次に、学習コーパスから計算された辞書や遷移確率などを用いて、テストコーパスの辞書検索や形態素解析の実験を行なった。表 1はそれぞれの方法における所要時間を測定した結果である。ただし、データの主記憶への読み込みにかかる時間を除いて計算してある。この時間は、オートマトンの記述を二次記憶に残す場合は 1 秒弱であり、主記憶に読み込む場合は 39 秒であった。表 1から DFA 法は AC 方に比較して辞書検索のみの場合で 11.7 倍、最尤経路の探索も含めた場合で 1.1 倍の速度であることがわかる。この値は、最尤経路の探索において浮動少数演算を避けたり、簡略化した未知語モデルを用いた場合大きくなると考えられる。

表 1: 辞書検索と形態素解析の速度 (文字 / 秒)

	AC 法 (二次記憶)	DFA 法 (二次記憶)	DFA 法 (主記憶)
辞書検索	220.7	2590.6	73242.1
形態素解析	94.9	108.2	120.0

以上の結果から分かるように、 AC 法と DFA 法では計算時間と記憶域のトレードオフがあるので、形態素解析システムの実装には、語彙数と計算機性能を考慮してどちらの方法を採用するかを決定すべきである。 AC 法や DFA 法などの高速の辞書検索方法を用いた結果、形態素解析全体の計算時間に占める辞書検索の時間は大きくなくなるので、形態素解析の時間的性能は最尤経路の探索のアルゴリズムや未知語モデルに大きく依存することが実験的に示された。オートマトンを主記憶に載せる場合の DFA 法は現状のハードウェアでは最も速い辞書の実装法であると考えられるので、形態素解析をさらに高速化するためには、解の探索と辞書検索と未知語モデルを一括して考える必要があると考えられる。

5 おわりに

本論文では、形態素解析の辞書検索を決定性オートマトンを用いて実現する方法を提案し、その時間的・空間的性能を AC 法と比較しながら論じた。決定性オートマトンを用いる方法は最も高速であることが理論的に示されるので、辞書検索を独立に考える限り、本論文で提案した方法が最も高速である。

今後、形態素解析をさらに高速化するためには、解の探索と辞書検索と未知語モデルを一括して考える必要がある。これは、確率オートマトンの最尤経路の計算がある種の半環上に成分をもつ行列の積の計算と等価であることを用いて実現されると考えられる。

謝辞

本研究を行なう上で、日本 IBM 東京基礎研究所の丸山宏氏に有益な資料や示唆に富むコメントを頂きました。ここに、感謝の意を表します。

参考文献

[1] 長尾真, 佐藤理史, 黒橋禎夫, 角田達彦. 自然言語処理. 岩波講座ソフトウェア科学 15. 岩波書店, 1996.

[2] 青江順一. ダブル配列による高速ディジタル検索アルゴリズム. 電子情報通信学会論文誌, Vol. J71-D, No. 9, pp. 1592–1600, 1988.

[3] 颯々野学, 難波功. 利用者による調節が可能な高速日本語形態素解析. 情報処理学会第 52 回全国大会, 第 2 巻, 1996.

[4] Hiroshi Maruyama. Backtracking-Free Dictionary Access Method for Japanese Morphological Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 208–213, 1994.

[5] Alfred V. Aho. Algorithms for Finding Patterns in Strings. In *Handbook of Theoretical Computer Science*, Vol. A: Algorithms and Complexity, pp. 273–278. Elseveir Science Publishers, 1990.

[6] Ronald L. Rivest. On The Worst-Case Behavior of String-Searching Algorithms. *SIAM J. Comput.*, Vol. 6, No. 4, pp. 669–674, 1977.

[7] 石畑清. アルゴリズムとデータ構造. 岩波講座ソフトウェア科学 3. 岩波書店, 1989.

[8] 有川節夫, 篠原武. 文字列パターン照合アルゴリズム. コンピュータソフトウェア, Vol. 4, No. 2, pp. 2–23, 1987.

[9] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.