1F-5

多重照合型形態素抽出方式に関する検討

福島　俊一 • 菊地　芳秀 • 大山　裕 • 宮井　均
(　日本電気株式会社　C＆Cシステム研究所　)

1　はじめに

形態素抽出処理は、単語辞書を検索することによって、入力されたテキストに出現した可能性のある全単語（形態素）を抽出する処理である。文章解析の最初に不可欠な処理である上に、処理時間に占める比率が大きい（形態素解析では処理時間の7割を越える）ことから、高速化が強く望まれる。

本稿では、ハードウェアによる並列処理を想定した高速化方式の１つとして、テキストと単語辞書との照合の際に、抽出位置や照合対象単語を多重化することを検討し、その実現上の問題点を明らかにする。

2　多重照合型形態素抽出モデル

多重照合型の形態素抽出モデルとしては、次の２通りが考えられる。

Ⓐ照合対象単語多重化モデル
Ⓑ抽出位置多重化モデル

従来のソフトウェアによる形態素抽出処理では、テキストのある位置から部分文字列を切り出して、その部分文字列に一致する単語を、単語辞書から検索する処理が繰り返される。その辞書検索の際には、部分文字列と単語との照合が、辞書内の単語ごとに逐次行なわれる。

これに対して、モデルⒶでは、テキストのある位置から切り出した部分文字列と、単語辞書内の複数の単語とを同時に照合する（図1参照）。このモデルⒶは、従来、連想メモリ[1]を用いて実現されている。ただし、現状の連想メモリLSIは、1個の容量が数k～数十kビットで、数万見出しの単語辞書をそのまま登録することはできない。価格性能比・規模の点から多数個の使用は現実的でないので、複数回に分けて登録・検索を繰り返す必要がある。

モデルⒷは、テキストにおける形態素抽出を行なう位置を多重化したものである（図2参照）。すなわち、テキストの1文字目・2文字目・3文字目・……から同時に抽出処理を行なう。このモデルⒷは、従来、テキストの各文字位置にPU（処理ユニット）を対応付けたマルチプロセッサ形式で実現されている[2]。この場合、単語辞書は共有メモリに置かれ、複数PUからのアクセス競合の調整機構が設けられる。

3　ISSPを用いた形態素抽出方式

ISSP（文字列検索LSI）[3]は、1個にL(=8)文字以下の単位文字列をN(=64)個まで登録できる。そして、入力される文字列に対して、登録された複数の単位文字列と並列に照合を行なうことができる。

以下では、ISSPを用いて、2で述べた2通りの多重照合型形態素抽出モデルを実現する方法を示す。

方式(1)はモデルⒶを実現する。従来の連想メモリでは扱えるデータが固定長のものが多いが、ISSPでは単語の表記のような可変長データを登録できる。ただし、登録・検索の繰り返しを必要とするのは同様である。

方式(2)と(3)は、テキストの部分文字列をISSPに登録して、検索時に単語辞書を入力することでモデルⒷを実現する。全抽出位置から辞書の同一部分をアクセスすることになるので、アクセス競合の調整機構は不要になる。

なお、以下の方式では、ISSPをS個×C個のマトリックス状に配置して使用する。同一列のS個は検索対象文字列が同一で、同一行のC個は登録内容が同一とする。

(1)単語辞書登録方式：　単語辞書の見出し部分をISSPに登録して、テキストのノンアンカー検索（任意の位置からの検索）を行なう。ただし、一度に登録できる見出し数はSN個までであるから、単語辞書をSN個ずつの部分辞書に分割して、部分辞書ごとの登録と検索を繰り返す。

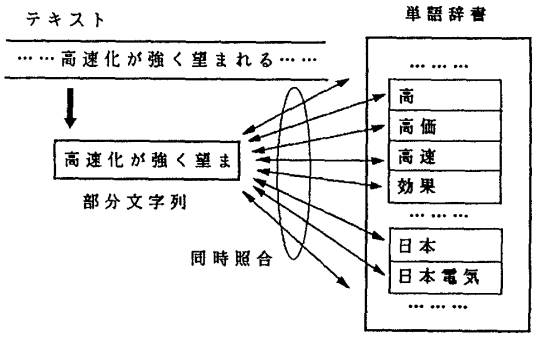

図1　照合対象単語多重化モデル

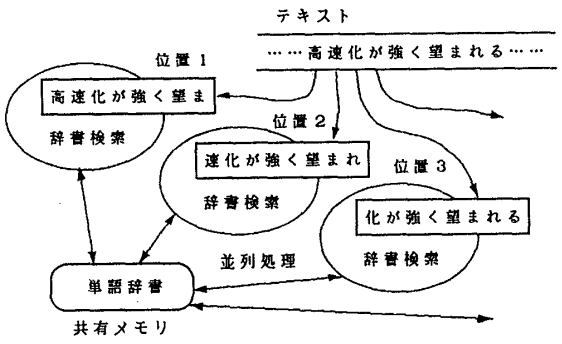

図2　抽出位置多重化モデル

A Study of the Morpheme Extraction Methods with Multi-Matching Technique
Toshikazu FUKUSHIMA, Yoshihide KIKUCHI, Yutaka OHYAMA and Hitoshi MIYAI
C&C Systems Research Laboratories, NEC Corporation

(2)ピラミッド型テキスト登録方式： ピラミッド型に展開したテキストをISSPに登録して、単語辞書の見出し部分のアンカー検索（デリミタで区切られた単位ごとの検索）を行なう。ピラミッド型に展開したテキストとは、同一文字位置から始まる文字列を、最大Lの異なる長さの文字列に展開したものの集合である（図3(b)参照）。単語辞書の見出し部分は、各見出しをデリミタで区切ってべた詰めした形式をとる（図3(a)参照）。

(3)シフト型テキスト登録方式： シフト型に展開したテキストをISSPに登録して、単語辞書の見出し部分のアンカー検索を行なう。シフト型に展開したテキストとは、文字位置を1文字ずつずらして長さLずつ取り出した部分文字列の集合である（図4(b)参照）。単語辞書の見出し部分は、各見出しに長さLまでワイルドカードを付加したものを、デリミタで区切ってべた詰めした形式をとる（図4(a)参照）。

ここで、単語辞書の総見出し数をD(=10^5)[個]、平均見出し長をW(=2.6)[文字]、ISSPの登録速度をE(=10^5)[文字/秒]、検索速度をV(=3×10^6)[文字/秒]として、長さX[文字]のテキストに対する形態素抽出処理時間T[秒]を求めると下記の式のようになる（式において、【x】はn-1<x≦nとなるような整数nを表わす）。値を代入し、S、C、Xに対するTのグラフを描いたものが図5である。

(1) T=[WN/E+(【X/C】+L-1)/V]【D/SN】
(2) T=[N(L+1)/2E+【D/C】(W+1)/V]
【[(X-L)L+L(L+1)/2]/SN】
(3) T=[NL/E+【D/C】(L+1)/V]【X/SN】

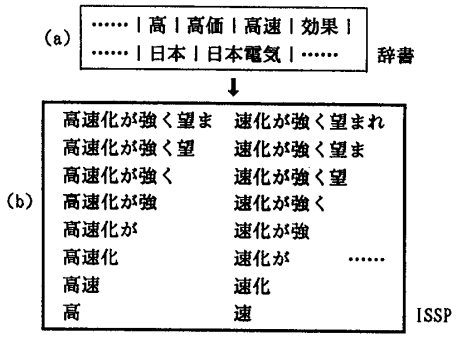

図3 ピラミッド型テキスト登録方式

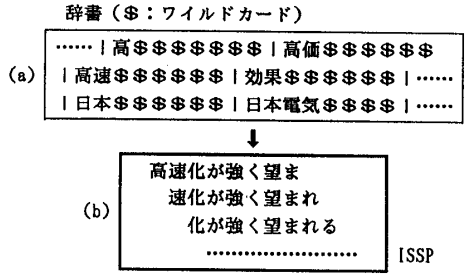

図4 シフト型テキスト登録方式

4 多重照合型形態素抽出方式の問題点

図5から、方式(1)は、単語辞書の繰り返し登録の時間が障壁となり、多量のテキストを一度に処理する場合でないと、方式(2)(3)に優る処理速度が得られないことがわかる。これは、連想メモリを用いた場合でも同様で、現状では、モデルⒶに共通の問題と思われる。

モデルⒷについては、マルチプロセッサ形式で実現すると、PU数を増加させても、アクセス競合により処理速度の向上は頭打ちになる。それに対して、ISSPを用いた方式(2)(3)では、アクセス競合が発生しないので、ISSP数を増すほど高速になる。特に、方式(3)の方が、ISSP数に対する処理速度が優れている。

しかし、方式(3)でも、図5の処理速度では、価格性能比・規模の面から実用化が難しい思われる。従来のソフトウェアでも、PC-98XL²(CPU:80386、ｸﾛｯｸ:16MHz)を用い、見出しを木構造化した単語辞書をRAMに置いて桁探索を行なえば、1秒間にテキスト100文字程度の形態素抽出処理速度が得られる。これより1桁以上高速な処理速度を想定すると、方式(3)でもISSPが6個以上必要になる。

さらに、モデルⒷでは、抽出される形態素の順番がテキスト内位置についてランダムになり、引き続いて行なわれる接続検定処理との整合が悪いという問題もある。

5 おわりに

多重照合型形態素抽出方式の問題点を明らかにした。

形態素抽出処理のハードウェア化には、多重化よりも、基本的な1回の照合処理のサイクルを可能な限り短縮するアプローチ[4]の方が適していると思われる。

参考文献

[1]小倉・他、連想ﾒﾓﾘLSIの現状と今後、信学誌69(7)、1986。 [2]中村・他、形態素抽出ｱﾙｺﾞﾘｽﾞﾑの高速処理方式、情処37全大、1988。 [3]山田・他、文字列検索LSI、信学技報CAS87-25、1987。 [4]福島・他、文章解析ｱｸｾﾗﾚｰﾀ(1)ー形態素抽出ﾏｼﾝの試作ー、本予稿集、1989。

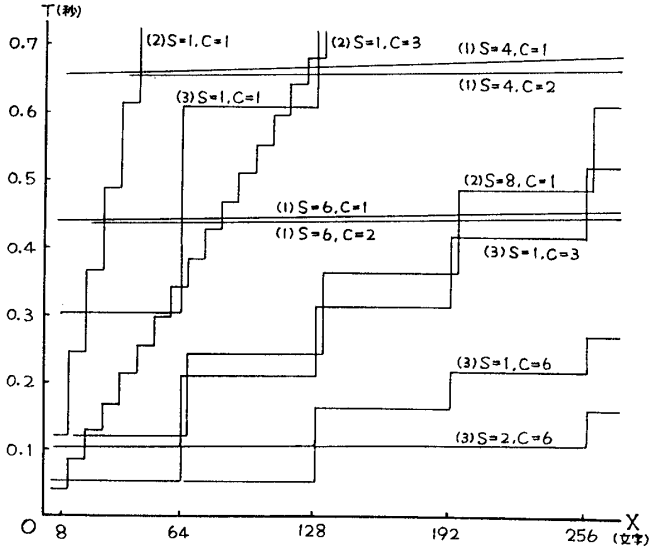

図5 ISSPを用いた形態素抽出処理時間