日本語情報処理用の計算機システム

6H−3

黒川一夫，大須賀勝美

東京理科大学

１．はじめに

現在の計算機は欧米で開発され発達してきたものであり、その根底は欧米の言語体系に基づいている。欧米語は表音文字であるアルファベットを用いているため、表意文字である漢字を用いている日本語とは大きく異なり、日本語を計算機上で扱うのは大変である。また、現在の計算機は数値計算を行うことを目的としたものであり、言語処理を行うのには必ずしも適したものとはいえない。計算のようにあらかじめ作業の手順が決まっている場合、その手順をプログラムとして書き表し、それに基づいて実行がなされる。そこで、現在の計算機のようなストアードプログラム方式が生まれた。しかし、言語情報の処理を行うには必ずしも手順が決まっているとは限らず、作業は逐次的で結果によって次の作業がいろいろと決定し、それを実行して行くこともある。したがって、言語の処理、特に日本語の処理を行うためには、従来のものとは異なる、日本語の言語体系に即した言語処理専用の新しいアーキテクチャを持つ計算機の登場が望まれる。そこで、当研究では日本語の特徴を生かした情報処理用計算機のアーキテクチャについて検討を行っている。

２．日本語の特徴

2.1 漢字の使用

日本語の最大の特徴は表意文字である漢字を用いていることであり、ここが表音文字を用いている欧米語と大きく異なる点である。漢字の使用により日本語で扱われる文字数は非常に多く、そのため計算機上で漢字を使用できるようになるまでに多くの時間を要した。逆に、欧米語のアルファベットは文字数が少なく、これらの組み合わせだけで全ての言葉が作られるため、計算機に早くから導入することができた。漢字は表意文字であるために文字それ自身が意味を持っており、１つの文字が表す情報量は非常に多い。また、文字が意味を持っているために文字同士のつながり方にも制限があり、文字数は多いが熟語を構成することのできる文字の並び方は限定されている。したがって、漢字を含む文章の文字列のパターンには特徴があり、そのパターンによって熟語を抜き出すことができる。また、漢字の意味を表す情報としてその漢字の部首が重要な役割を果たし、部首によりその漢字の特性を知ることができる。

2.2 熟語の構成

文字がいくつか並んで熟語が作られ、更にその熟語が集まって文章は構成されている。したがって、文章を熟語単位に切り分けることにより、文章を各熟語とそれらの並びとして取り扱うことができる。この際、熟語中での文字の並び方は正確に決まっており、ある順番に文字が並ぶことによって熟語となる。しかし、熟語の並び方については日本語の場合には語順がある程度自由であるから、文章の内容は熟語の並びでなく、その組み合わせだけによって決まる。

2.3 文字の種類

日本語では漢字の他に表音文字である仮名として平仮名とカタカナが存在するが、意味をもたないために文字の並び方にはあまり制限がなく、漢字のように熟語に切り分けるのは難しい。また、日本語の文章は欧米語のように単語の間に空白はないが、多種類の文字を使用しているため、そこから熟語の切れ目を知ることができる。

2.4 文章の語順

日本語の文章は文法的に語順に関する制約が少なく、比較的自由に語順を並べ替えることができる。文章の意味をとらえる際には、格助詞である「て」「に」「を」「は」が重要な役割をはたし、これらによって語順が入れ替わっても正しく文章の意味が理解できる。

2.5 意味の表現

漢字は文字それ自体が意味を持っているので、文字を並べて熟語を作れば、組み合わせられた漢字の意味から必然的に意味が生成される。日本語の場合には、文字の並び方によって意味を定義づけるような処理は必要なく、漢字は表意文字であるから文字を見ただけでその意味を推測することができる。したがって、文字の並びに対して意味を定義づける作業は、表音文字の言語処理においては必要であるが、日本語の言語処理においては無駄なことである。

３．日本語文章の取り扱い

日本語の文書データは、漢字などの文字がいくつか並んで単語や熟語が構成され、さらにそれらが集まって文章となり、文章が集まって１つの文書となる。

同じ分野の文書、特に専門分野の文書では、キーワードとなる同じ熟語が繰り返し使われるため、熟語の使用度は高くなり、同じパターンの情報を利用して文書情報を圧縮するとデータ量の縮小に効果がある。また、文章を理解するには、まず文章を解析して熟語単位に分解しておく必要があり、あらかじめそのような処理をしておくことにより検索などの効率が上がる。

Computer System for Japanese Language Processing.

Kazuo KUROKAWA, Katsumi OSUGA

Science University of Tokyo

漢字の取り扱いは、文字の意味を生かすために部首情報を用いてコード化し、すべての文字を部首コードとその中での順位という２つの情報で表現する。平仮名、カタカナ、英数字なども特殊な部首の文字として同じコード体系を用いる。

日本語の文書情報を熟語情報と結合情報の２つに分解して取り扱うことにより、文書の内容とその並び方の情報が得やすくなる。これにより、使用度の高い熟語情報が圧縮されてデータ量の縮小が可能になる。そこでデータを検索する際、いちいち長さの不定な文字列情報を取り扱わなくても、固定長の結合情報の中からある熟語番号の使われている場所を調べるだけで良くなる。

日本語の文章情報をここでは、以下のような５つの熟語情報に分類して取り扱うこととしている。(1)～(4)は直接的に日本語の文章として重要な情報であり、(5)は数式、数値などの計算処理に必要な情報である。

(1) 漢字だけの並びによる熟語
　（大部分が名詞で更に細かく分解可能）
(2) 漢字列の後に送りがなとして平仮名列が続くもの
　（ほとんどが動詞で活用あり）
(3) 平仮名だけの並びによるもの
　（助詞、助動詞、接続詞など）
(4) カタカナだけの並びによるもの
　（外来語を表しほとんどが名詞）
(5) その他の特殊な文字
　（数字、アルファベット、記号、など）

４．新しいアーキテクチャへの要求

言語処理、特に日本語情報処理の専用機として作業効率の良い計算機システムを目指すための、アーキテクチャへの要求を以下で検討する。

日本語処理の中心的な作業としては、自然言語からの情報の獲得、蓄積、引き出し、すなわち、データベースの作成や情報の検索といったことが考えられる。

熟語情報は文字の種類と文字数によって熟語を分類して整理しておくため、同一データ長のものをまとめて取り扱うことができる。そこで、データ単位に一括して処理が行えるようにするために、長いデータ単位で文字列の処理ができるようなメモリ構成を作る必要がある。

また、数式や数値の処理についても、従来の計算機とは異なった独自の方式をとり、演算装置を使用せずに、演算テーブルによる計算や公式などのパターンによる数式処理を行う。そして、プログラム化された処理ではなく、過去の知識を生かしていく、学習による処理方法を考える。

現在の計算機システムは数値計算用に考えられたものである。言語処理を扱う際にはメモリ構造において、次のような問題点が指摘される。

・同時にアクセスできるデータ長が限られ短い。
・メモリが１次元的なシリアルな配列である。
・同じメモリをプログラムやデータで共有する。

文書情報は前から後ろへという１本の流れでできている、シリアルな１次元的なデータである。したがって、プログラムや数値データと異なり、キャッシュメモリを使用した場合の効果は絶大である。また、文字列データは数値データなどと異なり、データ長が不変で長いため、一度にまとめてデータの取り扱いが行えない。

このようなことから日本語処理用の計算機として、その処理手順、データ構造、メモリの構築、において次のような新しいアーキテクチャが要求される。

・文章の情報は、熟語情報と結合情報の２つに分けて取り扱う。
・文字のコード体系を部首コード１バイト、部首内番号１バイトの２バイトコードで統一し、このコードによって熟語情報を組み立てる。
・部首コードから漢字、平仮名、カタカナ、数字などの文字の種類が分かるようにする。
・熟語情報の長いデータを１度に処理できるようにして、その長さも変えられるようにする。
・結合情報は種別、文字数、要素番号を含み、２バイトの固定長データとして扱う。
・メモリの構成として、プログラム領域とデータ領域とは分離させ、データ領域は固定長データを扱う部分と、可変長データを扱う部分とからなる。
・可変長のデータを一度に操作できるようなメモリ構造にする。
・各処理は、ある限定された作業領域を用いて分割的に作業を進め、この作業領域には高速のメモリを用いる。

５．まとめ

これまで、日本語処理用計算機の新しいアーキテクチャについて述べてきた。これは、日本語の情報処理を基礎としたシステムであり、その応用範囲としては文章の入力、文書の保存、文章・文字列の検索、情報の蓄積、などの作業が考えられる。さらにこのシステムを発展させることにより、将来的には文章の解析や理解、人工知能といったことへの展開も考えられ、これらの作業を従来の計算機システムで行うよりも効率的に処理が可能になるはずである。

＜参考文献＞

[1] 黒川一夫，大須賀勝美：
“日本語及び漢字を使用したコンピュータ”，
北京国際シミュレーション学会，(1989.10)．
[2] 黒川一夫，大須賀勝美：
“日本語を基礎とした計算機システム”，
信学技報　Vol.90,No.143,pp.35-40,(1990.7)．
[3] 黒川一夫，大須賀勝美：
“日本語を基礎とした言語処理用計算機システム”，
第９回　日本シミュレーション学会　論文集，
pp.239-242，(1990.6)．
[4] 高橋延匡：“日本語情報処理”，
近代科学社，(1986.7)．
[5] 長尾真：“日本語情報処理”，
電子通信学会　コロナ社，(1984.5)．