日本語処理用計算機のアーキテクチャ

6H−4

大須賀勝美，中川俊之，石松謙治，黒川一夫

東京理科大学

1．はじめに

現在の計算機は、日本語を取り扱うのにはあまり適当なものではなく、作業を行う上で困難なものとなっている。そこで、日本語の特徴を生かした新しい構造を持つ日本語処理専用の計算機システムの構築を目標とし、そのための新しい計算機のアーキテクチャについて検討を行う。

2．日本語処理用計算機のアーキテクチャ

日本語処理用の新しい計算機システムを実現するために要求される事項と、それを満足するアーキテクチャに関して以下に検討を行う。

(1) 日本語コード

現在の計算機システムでは、漢字や仮名といった日本語を取り扱うのに、便宜上通し番号をふって使用しているに過ぎず、表意文字としての情報を何も生かしてはいない。そこで、同じ２バイトの情報を使用して、１バイトで部首コード、もう１バイトで部首内の順位を表すようにして、漢字の部首情報を利用するようにする。また、平仮名やカタカナや英字、数字、記号といった文字は特殊な部首に属するものとして取り扱い、文章中で使われる文字はすべて２バイトのコードで表現する。熟語情報として文字の種類ごとに分類された後では、文字の区別はついているので、平仮名、カタカナ、数字、アルファベット、記号などは部首内順位の１バイトだけで表現できる。

(2) データ形式

日本語文章情報を熟語単位に分解して、熟語情報と結合情報という２つの情報に分けて取り扱う。熟語情報はその文書内で使用された熟語を集めて整理したものであり、文書の内容を表している。結合情報はその並び方を表したものであり、文章の流れを示している。文書ごとにこれら２つの情報を作成し、各文書ごとに矛盾のないデータとする必要があり、これらの情報から元の文書を復元することが可能でなければならない。

熟語情報は可変長の文字列データとして取り扱い、熟語の種類、文字数によって分類し、日本語コードで表現して部首コード順に整理しておく。結合情報は固定長のデータとし、使用された熟語の並び方を熟語の種類、文字数、熟語番号で記録している。

(3) メモリ構造

メモリの構造は、プロセッサ内部のレジスタ、キャッシュのような動作をする作業領域としての高速メモリ、作業中のデータを一時格納するテンポラリメモリ、データを保存するための外部メモリ、の４つの階層からなる構造とする。

プログラム領域とデータ領域に分離して情報は格納し、データ領域はさらに固定長と可変長の部分とに分かれる。

(4) メモリの管理

メモリを一定サイズごとのブロックに分け、ブロックごとにそのメモリ内を管理するプロセッサ部を設ける。このメモリブロックごとに、プログラムやデータを格納して、それぞれの領域を分離する。また、メモリブロックをいくつか組み合わせて同時にアクセスできるデータ長を自由に設定できるようにし、その長さでデータを扱い、比較、移動などの処理を行うようにする。

(5) メモリのキャッシュ化

文章の情報は基本的には一次元的なシリアル情報であり、このようなデータの処理に対してメモリをキャッシュ化する場合にそのヒット率は高く、そのメモリの高速化の効率は非常に高く、作業をうまく処理すれば１００％のヒット率も達成できる。

(6) メモリブロックの構成

・コンピュータの扱えるデータ長に依存せず、可変長のデータを１つの単位として処理することを目的とする。
・メインプロセッサは、いくつかのメモリブロックを組み合わせるが、組み合わされているメモリブロックの番号を管理している部分があり、それを経由してアクセスされるメモリブロックが指定される。

(7) メモリプロセッサの機能

・メモリプロセッサはある一定の大きさを持つメモリを管理し、アドレス情報やデータ情報の格納を行うレジスタを内部に持ち、メインとなるプロセッサとの間で情報の受け渡しを行う。
・管理するメモリ内のアドレスを指定してデータの読み書きを行い、データの移動、比較、入れ替えを行う。
・メモリ内のデータを一括して並び換えたり、メインプロセッサからデータを受けてメモリ内を検索する。

(8) プログラムの実行

数値計算の時のようにあらかじめ処理が決定していないので、完全にプログラム化することはせず、手順の決まっているいくつかの作業ごとにその処理についてプログラムを作っておく。そして、それらの組み合わせはメニュー形

Architecture of Japanese Language Processing Computer.

Katsumi OSUGA, Toshiyuki NAKAGAWA, Kenji ISHIMATSU, Kazuo KUROKAWA

Science University of Tokyo

式にして逐次的に実行し、実行結果を人間が判断して順次命令を与えて処理が進められるようにする。また、処理を実行する際には一定の作業領域を高速メモリ上にとり、限定された領域内でデータを分割処理することにより、実質的にメモリの高速化を図る。

(9) 高速処理化

文書情報のデータを処理する場合、その手順は数値計算に比べて簡単であるが、非常に大量のデータを取り扱う必要がある。そこで、データを分割していくつかのプロセッサで同時に処理を進める並列化が考えられる。また、データの読み込み、分類、ソート、圧縮、検索というような一連の手順をパイプライン処理して高速化することも考えられる。

(10) ソート・検索

言語処理を行う上で、ソートによるデータの並び換えと、その中からあるデータを探し出す検索といった作業は、使用頻度の高い処理である。そこで、あらかじめハードウェア的に処理が可能なようにしておけば作業効率は向上する。これらの処理はメモリ操作と密接に関係しているため、ブロックごとにメモリを管理するプロセッサ内部にその機能を持たせることにより、各メモリが独立しているので並列的に実行することも可能となる。

(11) 文章の入力

日本語文章の入力方法としては、既存の計算機で入力された文書情報をファイルから読み込む場合と、新しい計算機上で直接入力する場合が考えられる。

前者の場合はコード体系が異なるため、まずここで述べている日本語コードに変換し、その後に熟語情報と結合情報を作成して文章を表す。一方、後者は入力の段階ですでに日本語コード化されており、さらに熟語の切り分けが、かな漢字変換時に指定されてはっきりしているので、直接、熟語情報と結合情報という形でデータを格納していく。

3．アーキテクチャの評価

今回は、既存の計算機上でのシミュレーションにより、そのアーキテクチャの効果を調べて検討し、文書の変換や検索に対する評価を行っている。また、メモリの構造についても、その速度と効率の点についてもシミュレーションによりその効果を調べている。

検索処理などは実際の処理時間で比較するには、環境やプログラムなどの条件によって結果が異なるし、新しいアーキテクチャを持つ計算機も実在しないため、ここでは理論的に処理に要する時間を算出し、それに基づいて性能の比較を行っている。

実際に、いろいろな分野の文章について、文書情報を熟語情報と結合情報といった形で取り扱い、その評価を行った。代表的なものとして、教科書、専門書、用語辞典、法律文、哲学書についてデータをとって評価した結果を表１と表２に示す。ここで、圧縮率と高速化の値は、同じ文書のデータ量と検索時間を、従来の計算機上と新しいシステムの場合について比較したものである。この値が小さいほど効率は良いことになる。表１と表２から、圧縮率と高速化の結果はどちらも、文章の内容によらずほぼ同じような値が得られており、日本語文章ならほぼ全般的にこの取扱い方が有効であることがわかる。元の文書データに対する熟語情報と結合情報を合わせたデータの大きさを比較すると、データ量の圧縮は大体７～８割程度に縮小されることがわかる。検索の対象となる情報のデータ量を比べた検索の高速化に対する効率もほぼ４～５割と良くなり、空間的にも時間的にも効果的である。

表１　日本語文書データの圧縮効率

文書種別	ﾃﾞｰﾀ量	熟語数	使用度	圧縮率
電磁気学	155969	40893	2.24	0.67
制御工学	160260	36473	1.59	0.62
用語辞典	200819	41980	2.48	0.64
法 律 文	197716	44592	2.98	0.65
哲 学 書	188965	35018	2.41	0.67

表２　検索時の高速化の評価

文書種別	熟語の種類数	原文書検索量	変換後検索量	高速化
電磁気学	18240	78859	40893	0.52
制御工学	22998	80170	36473	0.45
用語辞典	16957	100797	41980	0.42
法 律 文	14976	98859	44592	0.45
哲 学 書	14535	94483	35018	0.37

また、実際にここで取り扱っている熟語情報と結合情報を利用し、文章の検索システムを構築して大きな文書中から文章を検索する場合も、目的とする文章を簡単に見つけ出すことができ、これらの情報でも検索が有効に機能していることが確認されている。

4．まとめ

現在は、日本語処理用の新しい計算機システムのアーキテクチャについて、シミュレーションにより検討を行っている段階である。将来的には具体的なシステムを構成して、実際の場合についての評価を行い、さらにシステムに改善を加えていく必要がある。

＜参考文献＞

[1] 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).
[2] 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).
[3] 高橋義造：“計算機方式”，
電子通信学会　コロナ社，(1985.7).
[4] 齊藤忠夫，発田弘：
“高性能コンピュータアーキテクチャ”，
丸善，(1989.3).
[5] 富田眞治，末吉敏則：“変列処理マシン”，
電子情報通信学会　オーム社，(1989.5).