4D−13

日本語情報処理用の計算機システム
―漢字処理機能の内蔵化―

大須賀 勝美，窪 敏喜，朝木 由貴，黒川 一夫
東京理科大学

１．はじめに

日本語の文章の中で、漢字は重要な役割を果たしており、日本語処理において漢字を処理する機能は中心的な位置を占めている。漢字は文字数が非常に多く、その取り扱いは欧米の文字などに比べて、はるかに困難なものとなっている。しかし、表意文字であるため、漢字同士のつながりに極めて制限がある。

本研究では、このような漢字の特徴を活かし、日本語処理に効果的な漢字処理方法を調べ、その機能を計算機システムに内蔵することを考える。

２．漢字処理機能

漢字の処理機能として考えられるのは、文章中での漢字文字列を単語あるいは熟語へと分解する作業と、かな漢字変換により文章を作成することなどである。

2.1 漢字文字列の分解

漢字は表意文字であるため、２文字以上の文字が並んで単語あるいは熟語となる組み合わせ方には制限が生じる。また、日本語の文章において漢字文字列は、ほとんどの場合が熟語を構成し、その熟語は更に漢字１文字あるいは２文字の単語へと分解が可能である。

日本語の文章を処理する場合、文章を各単語成分に分解する必要がある。この際に辞書からそれぞれの単語の有無を調べて分解することもできるが、ここでは今までに処理してきた文章から文字の結合パターンを蓄積して、その度合いを利用することを考える。漢字同士が並んで使用された回数を調べておき、この値の大きい文字の並びを単語と考えて分解する。

2.2 かな漢字変換

かな漢字変換は、与えられた読みを細かく切り分け、それぞれの読みに対応する熟語や単語を辞書の中から探し出し、適切な漢字の組み合わせを決定し、文章を作成していく。

３．辞書の役割

日本語で用いられる辞書としては、国語辞書と漢和辞典があるが、前者は単語や熟語が登録されており、読みによって該当する漢字の並べ方を調べ出すものであり、後者は漢字１文字ずつの読みとその漢字と組み合わせて熟語あるいは単語を構成する組み合わせ方の情報が登録されているものである。

漢和辞典としての読みと漢字の対応付けは不変的なものであり学習する必要のないものである。国語辞書的なものは、漢字の組み合わせ方であるのであまりにも対象範囲が広すぎ、学習により必要なものだけが抜き出されていた方が処理が容易になる。

４．システムの構成

漢字処理機能を組み込んだシステムの構成として、図１のようなものを考える。計算機システムの入出力部分に漢字処理部を配置し、この部分で独立的に漢字の処理を行い専用の辞書を内蔵して処理に利用する。

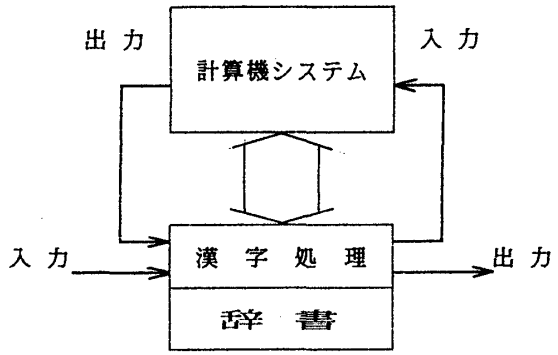

図１　漢字処理機能の内蔵

4.1 漢字処理部

文章情報は文字が並んでできた直列的な情報である。文字の並び方はシリアル・データとして入出力処理が行われる。しかし、単語や熟語は複数の文字が組み合わさって１つのデータとして意味を持ち、その処理には文字列長分の並列的な動作が要求される。

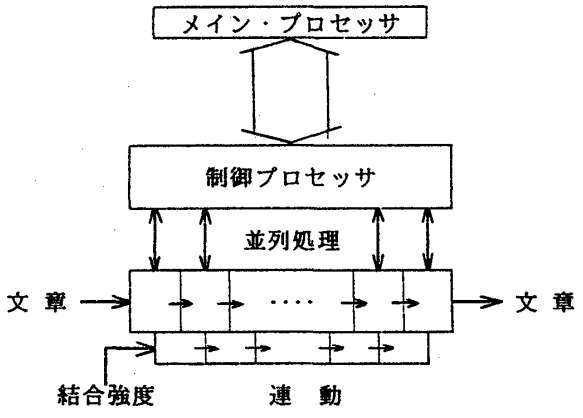

図２　漢字処理部の構成

Computer System for Japanese Language Processing
— Installation of Chinese Character Processor —
Katsumi OSUGA, Toshiki KUBO, Yuki ASAKI, Kazuo KUROKAWA
Science University of Tokyo

この作業を容易に行うために図２のような漢字処理部を考える。作業領域で並列的に動作して情報を集め、制御プロセッサが一括的に判断して熟語に切り分け、システムのメイン・プロセッサに結果を通知する。

4.2 漢字結合文字マトリクス

漢字同士の結合の度合いを調べて記憶する部分として、図３のような２次元マトリクス的な記憶装置を考える。図３の場合「電気」「磁気」「磁石」「電気」といったものの強度が強く、単語となることを示している。この情報は文字の並び方であり組み合わせではない。また、これは全漢字についてマトリクスが作成されるのでなく、使用されている漢字についてだけマトリクスが生成され、並んで使用された漢字間の結合強度が記憶させられる。

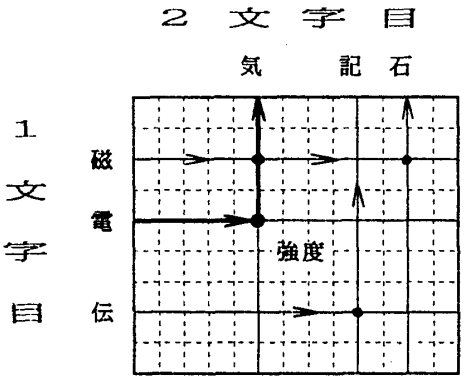

図３　漢字結合マトリクス

4.3 かな漢字変換処理

かな漢字変換処理としては、このシステムでは漢和辞典の役割を果たす、読み方に対する漢字だけを登録しておき、読みから考えられるいくつかの漢字の組み合わせの中でいちばん結合強度の強いもの結合マトリクスによって調べ、熟語あるいは単語として採用する。切れ方及び各文字の組み合わせの中から一番評価値の高いものを探し出す方法として、ニューラル・ネットワークのボルツマンマシンのような、並列的な動作ができるものを考える。

また、漢字からマトリクスの要素を結ぶ経路は連想メモリなどによって実現させ、読み入力から直接連想的にその読みをする漢字に対応する結合マトリクスへアクセスして、結合強度を調べられるようにする。

図３の例で「でんき」を変換する場合「でん」に対応する漢字が「電」と「伝」、「き」に対応する漢字が「気」と「記」でありこれらの結合強度を調べると「電気」「伝記」は強く「電記」「伝気」弱くなる。

4.4 結合強度の自動学習

結合強度としては文字と文字の並んだ頻度が記憶されているが、文章の中でその文字の並びが使用されると、図４のような構造により自動的に結合強度が変化させられ、知識を獲得していき、学習機能を果たすようにする。

このようなマトリクスの具体的なイメージとしては、１入力１出力のものではなく、同時多入力多出力のできるネットワークが必要であり、ニューロンのような細胞により各文字に１つずつ細胞を割り当て、使用された漢字の並び方により、その漢字を表す細胞間の結合強度を変化させていくような方法をとる。

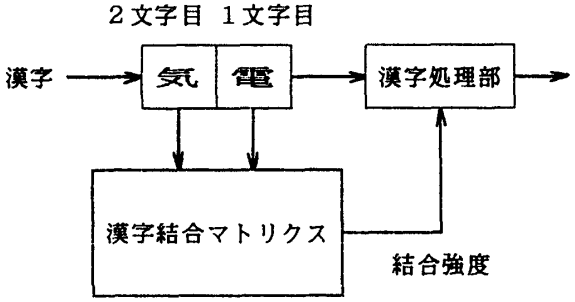

図４　結合強度の自動学習

また、辞書としての結合マトリクスの内容は、システムに固有のものであり、そのシステムで処理された文章によって内容が異なってくる。したがって、処理内容によって知識の専門化が図れる。幅広い分野の単語を覚えておく必要はなく、現在対象あるいはある個人が対象とする分野の単語だけが中心的に記憶されていく。また、辞書の内容は外部記憶装置へ書き出したり、外部装置から読み込むことを可能とすることにより、別のシステムで利用したり、複数の辞書の内容を統合整理することもできるようにする。そして、処理する内容に適した専門的な辞書を使い分けることにより、効果的な処理を実現する。

５．まとめ

文章の分解処理と作成処理について、並列的な動作を既存のコンピュータ上でシミュレーションし、その動作を調べる。この際、処理の対象となる文章中での文字の並びから、文字マトリクスの結合強度を学習させ、その後に、文章の分解やかな漢字変換による文章の作成を行わせる。これらの処理を行う際に、処理された情報に基づき結合マトリクスでの結合強度を変化させ、学習を行わせる。

最後に、ここでは漢字処理機能を計算機システムの中へ内蔵するための１つの方法を考えたが、今後更に、実際にかな漢字変換部を連想メモリで実現したり、結合強度の学習動作をニューラルネットなどで実現する方法を、具体的に検討していく。

参考文献

大須賀 勝美，黒川 一夫：
“日本語を基礎とした計算機システム”，
信学技報 Vol.90,No143,pp.35-40,(1990.7).
山田 八郎，本村 真人，榎本 忠儀：
“集積回路技術からみた機能メモリの現状と展望”，
情報処理 Vol.32,No12,pp.1239-1248,(1991.12).