文字列検索ＬＳＩを用いた国語辞書システムの構築法

5B-10

福島　俊一、　菊地　芳秀
(　日本電気株式会社　C&Cシステム研究所)

1　はじめに

電子化された国語辞書を人間の検索要求に応じて検索し、その検索結果を人間に理解可能な形式で提示するシステムを、ここでは、国語辞書システムという。国語辞書システムは、紙の国語辞書より柔軟で高速な検索が可能なため、単体システムとして有用であるばかりでなく、ワードプロセッサと統合化して、非常に効率のよい文章作成支援環境[1]を形成し得る。

以下、本報告では、国語辞書システムの一構築法として、文字列検索ＬＳＩ（ＩＳＳＰ：Intelligent String Search Processor）[3]を用いた、国語辞書全文検索方式のアプローチを示す。

2　全文検索方式のアプローチ

国語辞書には、各見出し語について、読み・表記、文法情報、語釈、関連語情報などが記述されている。国語辞書システム内でのそれら表現形式は、(A)紙の国語辞書の記述法そのままと、(B)概念記号やリンクを用いて抽象化・構造化した表現法とが考えられる。

(B)の形式は、辞書の作成に多大な経費と労力を要する上に、辞書知識の解釈方法がデータ構造によって固定されてしまうという問題がある[4,5]。そのため、(B)の形式の辞書は、機械翻訳など言語解析プログラムの参照する辞書としては必要であるが、人間が検索する国語辞書システムで用いるには実用的でない。

これに対して、(A)の形式では、辞書知識の柔軟な解釈・利用が可能である。しかし、従来、(A)の形式の辞書は、検索にインデックスを用いているので、あらかじめ設けたインデックスによって、検索方法が限定されてしまっている。しかも、人間向けの記述は、種々のゆれや曖昧さ（例えば、項目の順序の例外、関係を表現する語の使い方の不統一、語釈文に埋もれた異表記・関連語など）を含んでいるため、そこから、あらかじめ正確に情報を抽出しておくことは難しく[6]、通常、明示された読みや正表記による検索が行なえるのみである（記述内容に関する検索を実現する際には、そのインデックスの作成に、大きな労力を要する）。

そこで、筆者らは、(A)の形式の辞書に記述された知識を最大限に利用するために、国語辞書全文検索方式を提案する。インデックスを用いず、国語辞書の全文を検索することによって、検索方法が柔軟になる。

3　ＩＳＳＰを用いたシステム構成

国語辞書全文検索方式の問題点は、辞書検索に時間がかかるため、検索要求が出されてから検索結果が提示されるまでの応答時間が遅いことである。しかし、この問題点は、ＩＳＳＰを用いることで解決できる。

ＩＳＳＰは、当社で開発された文字列検索ＬＳＩである。ＩＳＳＰ１個には、基本的には、８文字（１文字：１６ビット）以下のキー文字列を６４個まで登録できる。そして、ＩＳＳＰに入力される文字列に対して、登録された複数のキー文字列と、並列に照合を行なうことができる。その際の最大入力速度は、１０Ｍ文字／秒と高速である。さらに、１文字の異なり（誤り・欠け・混入）を許容した検索、ワイルドカード（任意の文字・文字列と一致するように定めた特殊文字）を導入した検索なども可能である[3]。

図１に、ＩＳＳＰを用いた国語辞書システムの構成を示す。検索プログラムは、キーボードから入力された検索要求に応じて、ＩＳＳＰにキー文字列を登録する。それから、国語辞書の全文をＩＳＳＰに入力し、

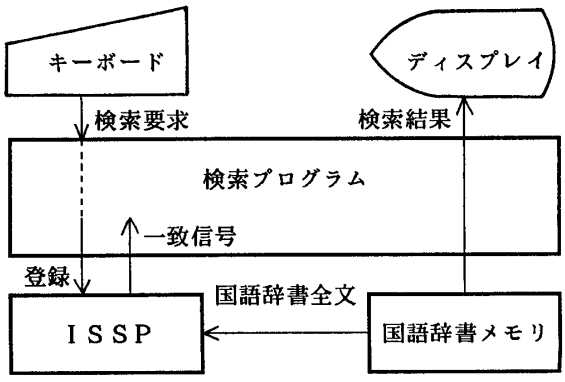

図１　国語辞書システムの構成

An ISSP Based Approach to Construct Dictionary Systems
Toshikazu FUKUSHIMA and Yoshihide KIKUCHI
C&C Systems Research Laboratories, NEC Corporation

ＩＳＳＰから得られる一致信号をもとに、国語辞書の、検索要求に該当する部分を、検索結果としてディスプレイに表示する。

なお、国語辞書を複数部分に分割し、その各々にＩＳＳＰを１個ずつ割り当てて、複数部分を並列に検索すれば、応答速度をさらに高速にすることができる。

4　国語辞書システムの機能

国語辞書の記述において、各見出し語の読みと正表記は明示されている（以下、読みは＜＞で、正表記は［］ではさまれて明示されているものとする）。そこで、国語辞書の文章を、読み・正表記・語釈文（読みと正表記以外の部分）の３項目に分けて、ＩＳＳＰを利用した検索機能を考える。

まず、読み・正表記の項目を対象に、完全一致検索、部分一致検索、曖昧検索の３機能が実現できる。これらは、キーボードから入力される検索要求の文字列と、該当項目（読みまたは正表記）の文字列との照合方法によって区別される。図２に、これらの検索機能の例を示す。図２の部分一致検索の例で用いている＊は、任意の文字列と一致するワイルドカードである。ＩＳＳＰはワイルドカードを考慮した照合が行なえるので、検索要求の文字列と、該当項目の文字列とが完全に一致する見出し語を検索するだけでなく、先頭・末尾など部分的に一致する見出し語を検索することができる。曖昧検索は、検索要求の文字列と、該当項目の文字列が１文字だけ異なる見出し語を検索する機能

	検索要求	検索結果
完全一致	＜しりつ＞	＜しりつ＞［私立］ ＜しりつ＞［市立］
	［市立］	＜いちりつ＞［市立］ ＜しりつ＞［市立］
部分一致	［私＊］	＜ししょばこ＞［私書箱］ ＜しせいかつ＞［私生活］ ＜しりつ＞［私立］　など
	［＊箱］	＜ししょばこ＞［私書箱］ ＜ふでばこ＞［筆箱］　など
	＜し＊つ＞	＜しさつ＞［視察］ ＜しせいかつ＞［私生活］ ＜しりつ＞［市立］　など
曖昧	＜メイデー＞	＜メーデー＞ ＜イデー＞

図２　読み・正表記の検索機能の例

であり、ＩＳＳＰは、完全一致検索と同等の速度で実行できる。この曖昧検索は、うろ覚えの語や、表記にゆれのある語の検索に有効である。

次に、語釈文を対象に、語釈文がある条件を満たす見出し語を検索する機能が実現できる。この機能は、ある意味をもつ語や、ある語の関連語（上位語・下位語・類義語など）を捜すために利用できる。例えば、条件の指定方法として、キーワードの論理式表現を考えた場合、「（昔∨過去）∧思い出す」という検索要求（∨：OR、∧：AND）から、［回顧］［回想］［旧懐］などが得られると期待される。このとき、活用するキーワードについては、「思い出｜す」のように、語幹と語尾との間に、あらかじめ定めた区切り記号が挿入された場合は、「思い出さ∨思い出し∨思い出す∨思い出せ∨思い出そ」のように、自動的に活用形の展開を行なうことも考えられる。

語釈文における語の関係や意味の記述方法にはゆれがあるため、語釈文の検索に試行錯誤的な一面があることは否めない。そこで、利用者が、使用してゆく間に、語の関係・意味を明示する記述を追加できるようにする。また、「の一つ」「の一種」など語の関係を規定する表現[6]を整理して、マクロとして用意しておくのも有効である。

5　おわりに

文字列検索ＬＳＩを用いて、全文検索を行なう国語辞書システムの構築法を示した。今後、ＰＣ－９８０１ＶＭ上に構築したテキスト検索システム[2]を用いて、方式の確認・評価を行なう予定である。

なお、このような国語辞書システムは、文章作成に役立つだけでなく、言語処理研究のツールとしても重要であると考える。

参考文献

[1] 福島・大竹・大山・首藤、日本語文章作成支援システムＣＯＭＥＴ、信学技報 OS86-21、1986
[2] 菊地・宮井、ＩＳＳＰを用いたテキスト検索システムの試作－論理構造単位の検索を中心に－、情処35全大 5S-9、1987
[3] 山田・平田・永井・高橋、文字列検索ＬＳＩ、信学技報 CAS87-25、1987
[4] 吉田、辞書構築における諸問題、情報処理 27(8)、1986
[5] 磯田・相磯・上林、辞書知識ベースのためのモデル、コンピュータソフトウェア 5(1)、1988
[6] 鶴丸・日高・吉田、単語間の上位－下位関係の自動抽出、情処研報 FI-3-1、1986