柔軟な検索機能を持つＫＷＩＣ検索システムの一方式

三吉秀夫　小渕保司　濱田明　秋山広勝
（シャープ株式会社 情報システム研究所）

1．はじめに

KWIC (Key Word In Context)は単語の用例を提供するためのデータベースであり、言語現象を分析したり自然言語処理システムの文法設計を行うために有用な言語ツールである[1]。しかし現存のKWICはその名の示すとおり、１個の単語の表記をキーとして検索するようになっているうえ、印刷物の形態をとるため必ずしもユーザのニーズに答えているとはいえない。自然言語処理研究者や言語学者が本当に調査したい言語現象は単なる１個の単語の用例ではなく、ある制約を持つ語句同士の共起、あるいはあるパターンを含む文例などであろう。そのためにも従来のKWICとは異なる文章検索システムの開発が望まれる。

我々はこれらの点を考慮し、柔軟な検索機能を持つKWIC検索システムの開発を目指している。本稿ではその概要を報告する。

2．検索機能の拡張

従来のKWICは単語の表記（見出し、代表形）をキーにして検索するようになっているが、本システムでは次のような拡張を行った検索キーワードを用いて検索を行うことができる。

①複数個のキーワード（キーワード列）により検索することができる。
②各キーワードに対して各種の制約を課することができる。
③ワイルドカードを用いることができる。

本システムではこのような条件を満たすキーワード列を検索語彙情報列と呼び、次のように定義される。

［定義1.］
検索語彙情報列とは、語彙情報を要素とする順序付リストである。

［定義2.］
語彙情報とは語彙範疇を定義する文法素性（grammatical features）の部分集合である。文法素性は"素性名／素性値"対として表現される。

つまり、１個の語彙情報が１個のキーワードに対応する。定義2.は"統語範疇は素性の集合として定義される（category as feature set）"という単一化文法のアイデアに基づく。語彙情報を形成する文法素性としては、品詞、品詞細分類、表記、見出し（代表形）、活用情報（活用形、活用型）、（シソーラスコードのような）意味情報などを考えている。

以上のように定義される検索語彙情報列を用いると、次のようなパターンをキーにして検索を行うことが可能になる。

（例）

(2.1) ［｛品詞／名詞｝，
｛見出し／の｝，
｛品詞／名詞｝］

(2.2) ［｛品詞／名詞｝，
｛見出し／を｝，
｛品詞／名詞，種類／サ変名詞｝，
｛見出し／する｝］

(2.3) ［｛品詞／名詞，意味／魚類｝，
｛見出し／を｝，
｛見出し／焼く｝，
｛品詞／名詞，意味／物質｝］

(2.4) ［｛品詞／副詞｝，
＊，
｛見出し／検索｝，
｛見出し／する｝］

(2.1)は「～の…」というパターンを含む用例を検索するための検索語彙情報列、(2.2)は「問題を検討する」というようなサ変名詞の用例を検索するための検索語彙情報列、(2.3)は「さんまを焼く煙」のような関係節を含む用例を検索するための検索語彙情報列である。(2.4)はワイルドカード（＊）を用いた例であり、ある副詞が「検索する」というサ変動詞と共起するパターンを含む文を検索するための検索語彙情報列である。ワイルドカードは０個以上の語彙情報と

A KWIC Retrieval System with Flexible Key Words
Hideo MIYOSHI, Yasuji OBUCHI, Akira HAMADA, Hirokatsu AKIYAMA
SHARP Corporation

マッチする。

3．実験例

我々は本システムのプロトタイプを、ICOTで開発されたPSIマシン上にCIL[2]を用いて開発中である。現在、検索される文章データベースはPSIの主記憶上に1文がCILの1個のユニットクローズとして実装されている。各単語に関する語彙情報はCILの部分項として表現され、各文はそれらのリスト形式をもつ。図1に検索結果の例を示す。図1は、

［（品詞／名詞），
（品詞／助詞，種類／格助詞，表記／で）］

という検索語彙情報列を用いて検索した例である。

4．おわりに

我々は現在、2で述べた基本機能をもつ検索システムを開発している。今後取り組むべき課題としては次のようなものが挙げられる。(1)検索処理速度の向上 — 現在、シーケンシャルな検索をしているため検索速度が遅い。バッチ処理を想定しているので特に問題にならないが、今後リアルタイム検索を目指してデータ格納方式、検索方式を検討する。(2)検索機能の向上 — 今は語彙レベルの情報しか使えないが、統語的な情報も使えるようにすること。また、語彙調査のような統計処理機能を設けること。(3)言語データの整備。

以上のような点を重視し、真に実用に耐え得る言語ツールとすることを目指して開発を進めてゆく。なお本研究は第五世代コンピュータプロジェクトの一環としてICOTからの委託として行ったものである。

［謝辞］
有益な御意見を頂いたICOT第２研究室の内田室長、吉川室長代理、及び研究員諸氏に感謝致します。

［参考文献］

1.長尾編，講座現代の言語7，言語の機械処理，三省，1984.
2.K.Mukai et al, Complex Indeterminates in Prolog and its Application to Discourse Models, New Generation Computing, Vol.3, No.4, 1985.

kwic_output（表記情報／非整列）

#HDR0240000000000074
／Arg｜の／要素｜数｜と／，／Predicate｜で／指定・し｜た／述語｜の／引数
#HDR0240000000000092
1｜回｜の／コミティケーション｜で／，／bind／-／hook／
#HDR0240000000000108
X｜か／コード｜で，／その／コード｜長｜と／Length／
#HDR0240000000000115
ア／（／エリア／1／～／3／）／実例｜で／実行・さ｜れ｜た／。／
#HDR0240000000000120
osition｜番｜目／（／0／オリジン｜で／指示／）／の／要素｜を／指・す／ロケー
#HDR0240000000000125
ア／（／エリア／1／～／3／）／実例｜で／実行・さ｜れ｜た／。／
#HDR0240000000000135
osition｜番｜目／（／0／オリジン｜で／指示／）／より／Length｜個｜の／
#HDR0240000000000155
／言語／処理｜系／システム／・／プログラム｜で／内部｜的｜に／生成・する｜ため／，／
#HDR0240000000000185
2／ビット／符号／無・し／整数／Y｜で／整数／時間｜を／行・う／。／
#HDR0240000000002

図1．検索結果の例