

構造化キーワードを用いた用例検索システムの試作

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)

構造化された検索キーワードを用いた用例検索システムの試作について報告する。従来の固定化されたKWICを用いた検索システムという観点から眺めた場合、1個の単語の表記に基づいた検索しか行うことができないものが多かった。本システムでは複数の単語から構成される検索キーワードを用いて検索を行うことができる。また検索キーワードを構成する各単語に文法的な制約を課したり、ワイルドカードを用いることもできる。このように構造化された検索キーワードを導入することにより柔軟な用例検索を行うことが可能になり、言語研究あるいは自然言語処理研究の分野において有用な言語分析ツールとなることが期待される。

An Experimental Text Retrieval System based on Structured Key Words

Hideo MIYOSHI, Yasuji OBUCHI, Akira HAMADA, and Hirokatsu AKIYAMA
Information Systems Laboratories, SHARP Corporation
492, Minosho-cho, Yamato-Koriyama, Nara, 639-11, Japan

This paper presents an augmented text retrieval system based on structured keywords. Most of the traditional KWICs have a text retrieval method based on the spellings of only one word. In this system, the user can search the texts using multiple search keywords. Besides, the user can give the various grammatical constraints to the keywords or can use the keywords including a wild card. By introducing this type of structured keywords, this system will provide the researchers in linguistics and natural language processings with a flexible text retrieval method.

1. はじめに

自然言語処理システムの開発において、解析や生成に必要な文法を開発したり言語理解に必要な知識ベースを構築するためには、ある語句が実際に使用されている多くの用例を参照することが不可欠である。また言語学者が言語現象を分析する場合も同様である。従来このような目的のためには『文脈付見出語』と呼ばれているKWIC (Key Word In Context) がよく用いられてきた。なぜならKWICは単語の用例を提供するための言語データベースであり、言語研究あるいは自然言語処理研究を行うための有用な言語分析ツールであるとされてきたからである。しかし従来の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仕組みであったり、印刷物の形態をとっているものが多いため必ずしも上記の用途に合った有用な情報が得られるとはいえない。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個々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複雑なパターンを含む用例といったものであろう。従来の固定化されたKWICではこのような複雑なパターンによる用例検索は不可能であるといってよい。より柔軟な検索システムが望まれる。このような動機から、我々は構造化された検索キーワードを用いて柔軟に用例検索を行うことができるシステムの開発を行っている¹⁾。本稿では、今回試作したプロトタイプシステムについて報告する。

2. 用例検索システムの必要性

本節では用例検索システムという観点から従来のKWICの問題点を考察し、更に本システムとの比較を行う。

2.1 従来のKWIC

KWICは、見出し語をそれが用いられている前後の文脈を付けてソートした表であり、国内においても多くのKWICが開発されている²⁾³⁾⁴⁾。通常のKWICは1つの単語をキーにして分類を行った形式を持つものが多い。一方、このような単語単位でソートしたKWICのほかに、多少異なる形式を持つものとしては電総研で開発された「日本語品詞列集成」⁵⁾があげられる。日本語品詞列集成は品詞付き単語列を品詞でソートし、所定の長さの品詞列とその品詞列に該当する単語の並びに前後文脈をつけて収録したものである。もともとKWICは主に文献検索のために考案されたものであるが、一方では言語研究のため、ある

いは自然言語処理研究の分野で有効な補助ツールとしても使用することが可能である。その理由は、KWICは大量の文章データの中から特定の単語の用例を抽出するためのツールと考えることができるため、言語学者が自分の文法理論の正しさを検証したり、新たな理論を構築する際の実証データを容易に参照することができるからである。また自然言語処理の分野においては、文法を開発する過程において単語の実例の用例を数多く参照しなければ広い言語現象をカバーできる十分に体系化された文法規則を書くことはできない。これは自然言語処理システムにおける辞書を開発する場合にも同様である。ある語の文法的な性質・制約・振る舞い(格パターンや特定の意味的な要素との共起関係など)を記述するためには、その語が実際にどのような文脈で用いられているかを広く調査する必要があるからである。KWICはこのような調査を行うために有効な言語データベースである。

2.2 用例検索システムへの要求

次に用例検索システムという観点から従来のKWICについて考察してみよう。前述のように従来のKWICは1個の単語の表記、あるいは決められたパターンを検索キーとして用例を検索する仕組みになっている。このように固定化されたもので言語学者が望むような言語情報を効率良く提供できるのであろうか。また文法・辞書を開発するために必要な文法的制約が容易に得られるのであろうか。我々はこれらの点に関しては若干の疑問をもっている。例えば中野は、言語学者が言語研究を行うにあたって検索したい情報として次のような項目を挙げている⁶⁾。

- (1) ある語が使われているか。
- (2) ある語がどの作品のどのページ・どの行で使われているか。
- (3) ある語がどの作品のどのページ・どの行で・
どういう文脈の中で使われているか。
- (4) ある語がどの作品のどのページ・どの行で・
どういう文脈の中で・
どういう品詞情報列の中で使われているか。
- (5) ある意味の語を文脈つきで出せ。
- (6) 連用修飾を文脈つきで出せ。
- (7) 主語を文脈つきで出せ。
- (8) 「～の～する名詞」構文の用例を出せ。
- (9) ある語にかかる語を出せ。
- (10) 明治上期上流階級の女性の会話文において人
称代名詞を使った文を出せ。

- (11) ある短い語が長い語の構成要素になっている場合の、その長い語を出せ。
- (12) 語を構成するすべての音節の母音が[a]である語を出せ。
- (13) 語頭子音が[r]である和語を出せ。
- (14) 語尾の母音連続が[ei]である和語名詞を出せ
- (15) 5モーラの和語名詞を出せ。
- (16) 語の頻度数調査の結果を出せ。
- (17) 語種別・品詞別頻度調査の結果(のべ)を出せ。
- (18) (17)ののべ・異なりについて出せ。
- (19) 会話文における(18)の結果を出せ。
- (20) ある段落の(18)の調査結果を出せ。
- (21) どういう漢字がどれだけ用いられたか。
- (22) ある語の表記は何種類あるか。
- (23) 文型の頻度数調査。
- (24) 「である体」で書かれている文章を出せ。
- (25) 漢語詞で書かれている文章を出せ。
- (26) 擬古文で書かれている文章を出せ。
- (27) 明治後期の東京の下町を舞台にした作品リストを出せ。
- (28) 白樺派作家の作品リストを出せ。
- (29) 私小説リストを出せ。

これらの機能のうち(1)から(15)まではある特定のパターンを含む用例を検索する機能である。(16)から(23)までは統計的調査を行う機能である。また残りはその他の種類の検索を行う機能である。従来の固定化されたKWICではこのような柔軟な調査・検索を行うことは困難である。KWIC作成プログラムとは異なる新たなソフトウェアが必要となる。

2. 3 本システムの特長

我々が開発している「構造化キーワードを用いた用例検索システム」は、2. 2で述べた要求のサブセットを満たすべく、次のような特徴をもっている。

- (1) 複数の単語から構成されるパターンを検索キーワードとして日本語文の用例を検索することができる。さらに、検索キーワードの各単語に各種の文法的制約を課することができる。このような性質をもった検索キーワード列を構造化キーワードと呼ぶ。文法的制約としては、「品詞、品詞細分類、活用情報、見出し、出現形」のような形態的情報のみならず、概念の階層関係を表すソーラスの概念名を意味的制約とし

て与えることが可能である。

- (2) 検索された用例はユーザが指定する順序にソートされ、KWIC形式で表示される。また検索結果中の特定用例について更に詳細の文法情報を提供することが可能である。
- (3) ユーザフレンドリな構造化キーワード入力機能を持つ。
- (4) 構造化キーワードの一部に任意の単語列とマッチング可能なワイルドカードを用いることができる。

次節に本検索システムの構成とより詳しい機能について述べる。

3. 構造化キーワードを用いた用例検索システム

3. 1 システム構成

本システムのプロトタイプは、ICOTで開発された逐次型推論マシンPSI上にCIL⁷⁾およびESPを用いて開発されている。図3. 1に全体のシステム構成図を示す。図のように本システムは主に「キーワード入力」部、「検索」部、「検索結果出力」部、「テキストデータベース」、の4つのモジュールから構成される。「キーワード入力」部は検索のための構造化キーワードの入力を行うモジュールであり、キーワードの単語数や各単語の文法素性を入力する。「検索」部は与えられた構造化キーワードをCILの部分項リストに変換し、テキストデータベースの検索を行う。「検索結果出力」部は検索した用例をユーザが指定するモードに従ってKWIC形式で表示する。「テキストデータベース」は「検索」部によって検索される対象となる(付加情報付き)テキストが格納されているデータベースである。次項より詳細機能について述べる。

3. 2 テキストデータベース

3. 2. 1 素性集合として表現される単語

「テキストデータベース」は「検索」部によって検索される対象となるテキストが格納されているデータベースである。2. 3で述べたような柔軟な検索機能を実現するためには、データベースには原文の文字づらの情報だけでなく、「品詞、品詞細分類、活用情報、意味情報」などの付加しておく必要がある。そのほか「単語の切れ目、出典、文番号」などの情報も必要である。このような情報を簡便に記述するために本システムではテキストデータの各単語を“素性名/素

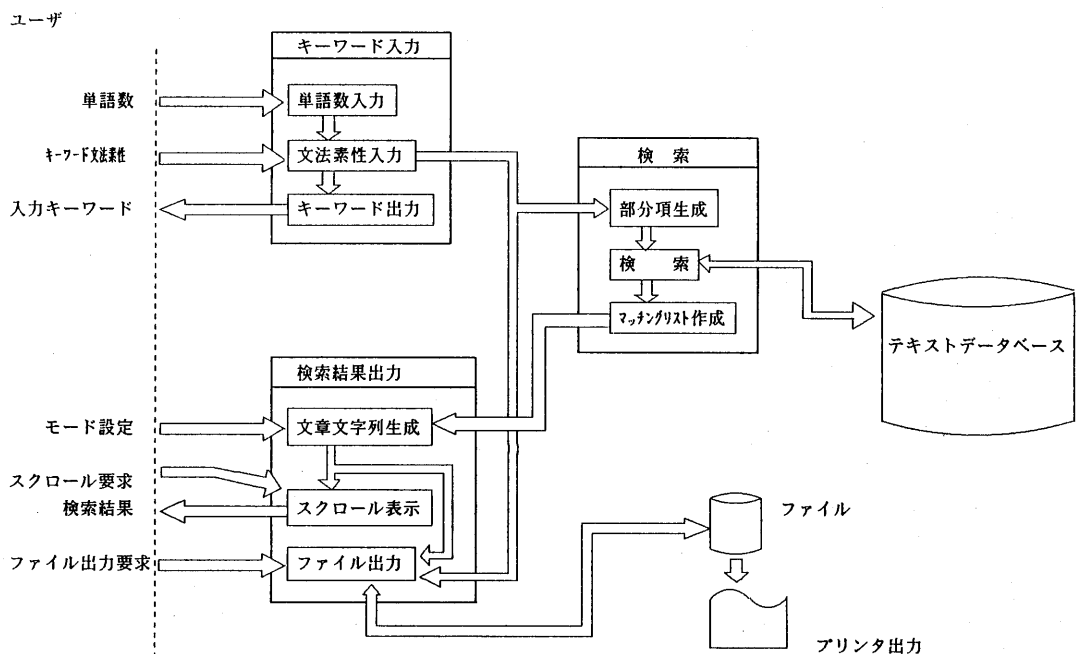


図3.1 システム構成図

性値” 対の集合としてとらえ、各種文法情報の集合として1個の単語を表現している。これは文法範疇を素性集合ととらえる単一化文法理論⁹⁾の枠組みに基づいている。そして各文の情報はその集合のリストとして表現される。各単語を構成する文法素性はその単語の品詞によって異なる。

3.2.2 品詞分類

表3.1に本システムで採用している品詞分類と各品詞が持つ文法素性を示す。表からわかるように、14種類の品詞を導入している。この品詞分類は一般的な学校文法に準拠している。ただし最後の「補助詞」は、一般に補助動詞、補助形容詞などと呼ばれているものの総称である。これらの各品詞に対して現在8種類の文法素性を割り当てている。ただし各品詞が8個すべての素性値を持つのではなく、表3.1の○印のついた素性のみを持つ。各文法素性の役割と値は次のとおりである。

品詞：14個の品詞名のいずれか。

表記：その単語が文中で実際に使用されている表記（出現形）。

見出し：活用語の終止形。複数表記のある場合は代表記。表記と同一の場合が多い。

種類：品詞によって異なる値を持つ。名詞の場合は

	品詞	表記	見出し	種類	活用	活用	活用	意味
動詞	○	○	○	-	○ (○)	○	-	
形容詞	○	○	○	-	-	-	○	-
副詞	○	○	○	-	-	-	-	-
名詞	○	○	○	○	-	-	-	○
連体詞	○	○	○	-	-	-	-	-
接続詞	○	○	○	-	-	-	-	-
代名詞	○	○	○	-	-	-	-	-
形容動詞	○	○	○	-	○	-	○	-
感動詞	○	○	○	-	-	-	-	-
接頭語	○	○	○	○	-	-	-	-
接尾語	○	○	○	○	-	-	-	-
助詞	○	○	○	○	-	-	-	-
助動詞	○	○	○	-	-	-	○	-
補助詞	○	○	○	-	-	-	○	-

表3.1 品詞分類と素性

「普通名詞、サ変名詞、数詞、固有名詞、形式名詞」のいずれか。接頭語および接尾語の場合は、それが接続する語によって「一般、数詞」

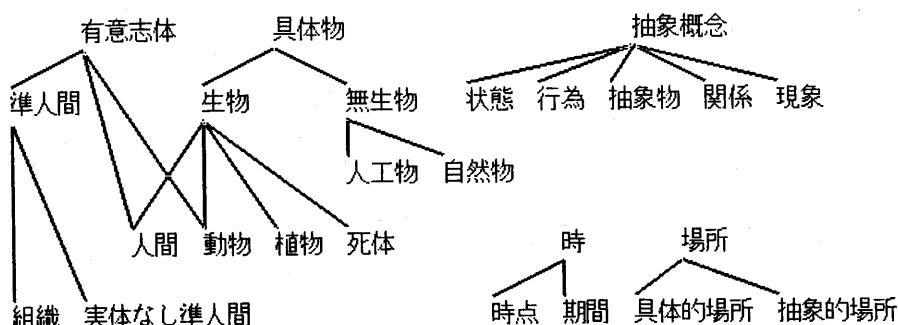


図3. 2 本システムで用いているシソーラス

のいずれか、助詞の場合は「格助詞、終助詞、副助詞、並立助詞、接続助詞、係助詞、準体助詞」のいずれか。

活用型：「五段、一段、カ変、サ変、行く、ある、なさる」のいずれか。

活用行：五段活用の動詞のみに付加する。

活用形：「未然、連用、終止、連体、假定、命令」のいずれか。

意味：図3. 2の本システム用シソーラスの中の概念名を要素とするリスト。複数個の値を取り得るのでリストにする。

3. 2. 3 文情報の実装

テキストデータベースのそれぞれの文は次のような形式をもつC I Lの1個のユニットクローズとして表現されている。

sentence(出典、文番号、文情報リスト)。

「文情報リスト」は各単語を素性集合として表現したもの(C I Lの部分項)を要素とするリストである。素性集合をC I Lの部分項として表現する理由は、部分項はフレームのようなデータ構造を表現し易く、C

I Lの拡張ユニフィケーションが使えるからである。この形式を用いることにより、例えば「自然を守る」という出典の13番目の文「森という環境を例にとってみよう」は図3. 3のように表現される。

3. 3 キーワード入力部

キーワード入力部は検索のための構造化キーワードの入力を行うモジュールであり、構造化キーワードの単語数や各単語の文法素性を入力する。

3. 3. 1 構造化キーワード

検索のための構造化キーワードは次のように定義される。

[定義3.3.1]

構造化キーワードは、語彙情報を要素とする順序付リストである。

[定義3.3.2]

語彙情報とは語彙範疇(単語)を定義するための文法素性の部分集合である。ただし文法素性は“素性名／素性値”対として表現される。

このようなキーワードを導入することにより、表記あるいは品詞のみをキーとしていた従来のK W I Cでは

sentence (自然を守る, 13,

[(品詞／名詞, 見出し／森, 表記／森, 種類／普通名詞, 意味／[具体的場所, 自然物]),
(品詞／助詞, 見出し／と, 表記／と, 種類／格助詞),
(品詞／動詞, 見出し／言う, 表記／いう, 活用型／五段, 活用行／ワ, 活用形／連体),
(品詞／名詞, 見出し／環境, 表記／環境, 種類／普通名詞, 意味／[状態, 具体的場所]),
(品詞／助詞, 見出し／を, 表記／を, 種類／格助詞),
(品詞／名詞, 見出し／例, 表記／例, 種類／普通名詞, 意味／[抽象物]),
(品詞／助詞, 見出し／に, 表記／に, 種類／格助詞),
(品詞／動詞, 見出し／とる, 表記／とつ, 活用型／五段, 活用行／ラ, 活用形／連用),
(品詞／助詞, 見出し／て, 表記／て, 種類／接続助詞),
(品詞／補助詞, 見出し／みる, 表記／み, 活用形／連用),
(品詞／助動詞, 見出し／よう, 表記／よう, 活用形／終止)]) 。

図3. 3 文情報の例

行えなかった柔軟な検索が可能になる。次に構造化キーワードの例を示す。

(例)

(3.3.1) [{品詞/名詞, 種類/普通名詞,
意味/抽象概念},
{品詞/助詞, 見出し/を}]

(3.3.2) [{品詞/副詞},
*,
{品詞/動詞, 活用型/五段,
活用形/連用}]

(3.3.1)は抽象概念という意味素性をもつ普通名詞の次に助詞「を」が用いられている用例を検索するための構造化キーワード、(3.3.2)はワイルドカード(*)を用いた例であり、ある副詞の後方に五段活用の動詞の連用形が現れる用例を検索するための構造化キーワードである。

3.3.2 構造化キーワード入力手順

構造化キーワードは次の順序で入力される。

- (1) 出典
- (2) 品詞
- (3) 表記・見出し
- (4) 種類
- (5) 活用型・活用行・活用形
- (6) 意味

素性値は、「表記・見出し」については直接入力するが、他の素性は専用ウインドウにおいてメニュー選択方式により行う。(4)～(6)の素性は品詞の値により不要な入力要求を出さないようにしている。構造化キーワード入力中の画面の例を図3.4に示す。

3.4 検索部

検索部は入力された構造化キーワードをパターンマッチ可能な部分項に変換し、テキストデータベースへの検索を行う。そして検索結果を蓄えておく。

3.5 検索結果出力部

検索結果出力部は検索した用例をユーザが指定するモードに従ってKWIC形式で表示する。ユーザは次のようなモード指定を行うことができる。

- (1) 表示する文法素性(デフォルトは表記のみ)
- (2) ソートするか否か(昇順, 降順, 出現順)
- (3) ソートの優先順位(どのキーワードのどの素性か)
- (4) 検索結果の出力先(CRT, ファイル, エディタのバッファ)
- (5) 詳細文法素性表示指定(特定検索文の詳細情報表示)

図3.5は(3.3.2)で示した構造化キーワードによって検索した結果の例である。本ウインドウは上下左右

のスクロールが可能である。

4. 今後の課題

今回試作したプロトタイプシステムは、小規模のデータベースながら言語研究あるいは自然言語処理研究のためのツールとして有用性のあることが確認された。今後の課題として次のような項目が挙げられる。

①検索効率の向上

現在のシステムはシーケンシャルな検索を行っているため検索速度があまり速くない。また、検索時間がデータベースの大きさに比例して増えてゆく。この点に関しては、バッチ処理を前提にしているので特に問題にはならないが、今後はリアルタイム検索を行えるような、データ格納形式、検索方式を検討する必要がある。そのためには処理の並列化、ICOTで開発されている分散知識ベース管理サブシステムKappa⁹⁾への移植なども検討したい。

②検索機能の向上

現在のシステムでは検索のための構造化キーワードとして単語単位の語彙レベルの情報しか与えられないが、次のような検索を行うことを可能にするため統語的(Syntactic)な情報も与えられるようにしたい。

- ・埋め込み文+格助詞「と」というパターンを含む文を出せ。
- ・関係節+名詞句というパターンを含む文を出せ。
- ・主語が欠けた文を出せ。

そのほか、現在非常に単純なシソーラスを用いているが、今後EDR等で開発されている¹⁰⁾¹¹⁾高度なシソーラスを参考にし、拡張の検討をしたい。

③言語データの整備と支援ツール

本システムを言語ツールとして実際に役に立つシステムとするには、テキストデータベースの大規模化を図り、出来るだけ多くの用例を検索できるようにする必要がある。おそらく最低10万例文程度のデータベースが必要であろう。そのためには、人手によって行っているデータベース作成作業を自動化する必要がある。ATRで開発されているような支援システム¹²⁾を開発する必要がある。

5. おわりに

本稿では、今回試作した“構造化キーワードを用いた用例検索システム”について報告した。今後試験的な使用を積み重ねながら改良を行い、真に実用的な言語分析ツールを目指してゆきたい。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からの委託により行っているものである。

key#1/2	%key%	2. key:動詞
品詞: 副詞	ワイルドカード	2. key:/
		2. 動詞:五段
		2. 動詞:指定なし
		2. 動詞:活用形
		cancel
		未然
		連用
		終止
		連体
		仮定
		命令
		指定なし
		フ
		ワ
		指定なし
		指定なし

図3. 4 構造化キーワード入力中の画面の例

kwic2:output (検索用例数:15)	
用例8	出典: 演講演2/第23文
現在ではむしろ一部の人が心配するようにAIブームは過	
用例9	出典: 演講演2/第35文
むしろこれまで得られたテクニックを使っていろいろ	
用例10	出典: 演講演2/第33文
は前からむしろAIの主流であったと思います	
用例11	出典: 演講演2/第35文
むしろこれまで得られたテクニックを使っていろいろ	
用例12	出典: 演講演2/第22文
が始まってもう4年たちましたしその前の準備段階の議	
用例13	出典: 演講演2/第22文
が始まってもう4年たちましたしその前の準備段階の議	
用例14	出典: 演講演2/第22文
からすればもう78年たっています	
用例15	出典: 演講演2/第26文
の見方をもう一度しっかりしなければいけないのではない	

図3. 5 検索結果の例

〔謝辞〕

本研究を行う機会を与えて頂きましたシャープ株式会社情報システム研究所の三坂所長、第2開発部の大崎部長に感謝致します。また本研究内容に関して日頃より御指導頂き、有益な御助言を頂きましたI C O T第2研究室の内田室長、吉岡室長代理、吉川前室長代理(現NTT)に感謝致します。最後に本システムのプログラム開発に御協力頂いたSBCソフトウェア株式会社の真田氏に感謝致します。

〔参考文献〕

- [1] 三吉, 小淵, 濱田, 秋山, 柔軟な検索機能を持つKWIC検索システムの一方式, 情報処理学会第36回全国大会, 4U-8, (1988).
- [2] 植村, 電子計算機による自動索引の研究(上), (下), 電子技術総合研究所研究報告第743号, 第747号, (1974).
- [3] 長尾編, 講座現代の言語7, 言語の機械処理, (三省堂, 1984).
- [4] 長尾監, 日本語情報処理, (電子通信学会, 1984).

- [5] 電子技術総合研究所編, 新編日本語品詞列集成, 右順編(1979), 左順編, (1979).
- [6] 中野, 言語研究用データにおけるデータベースの考え, 国立国語研究所内部資料, (1973).
- [7] K. Mukai and H. Yasukawa, Complex Indeterminate in Prolog and its Application to Discourse Models, New Generation Computing, Vol.3, No.4, (1985).
- [8] Shieber, S., An Introduction to Unification-Based Approaches to Grammar, CSLI Lecture Notes, NO.4, (1986).
- [9] 新世代コンピュータ技術開発機構, 電子計算機基礎技術開発成果報告書, 基礎ソフトウェアシステム編第1分冊, (1986).
- [10] 田中, 仁科, 上位/下位関係シソーラスISA MAPの作成[1], 同[2], 情報処理学会研究報告87-NL-64-4, 5, (1987).
- [11] 日本電子化辞書研究所, 概念辞書(第1版), EDR Technical Report TR-007, EDR, (1988).
- [12] 小倉, 篠崎, 森本, 言語データベース収集支援システム, 情報処理学会第36回全国大会, 4U-4, (1988).