解　説

情報検索のための日本語解析†

絹川　博之††

1. はじめに

漢字入出力機器の整備に伴ない，官公庁を中心に漢字かな混りの日本語文情報を対象とした情報検索システムのニーズが高まっている．例えば，大蔵省主計局[5]，外務省[4]，通産省[3]，国立国会図書館[3]，国文学研究資料館[9]などである．京都大学[1]，電子技術総合研究所[8]，日本情報処理開発協会[3]では，実験システムを開発しており，JICST[2]では，オンライン情報サービス JOIS-I の漢字化システムとして，JOIS-K を開発している．一方，民間では，日本経済新聞社が，記事検索システム[7]を開発し，サービスを開始している．このような状況を踏まえて，汎用日本語情報検索ソフトウェアの開発が，必要とされている．この時，単に漢字をサポートすれば，事足りる訳ではなく，使い易く効率よく運用できる機能が具備される必要があり，そのためには，日本語解析まで立ち入る必要がある．この課題に対処するために，日立では，記事検索における自動インデクシング[15]を具体的ケースとして，研究開発を進めている．以下，これを含めて，実用を目指した生データの解析について，方式，特徴，適用対象等について述べる．なお，富士通[5]では，大蔵省システムを例に，東芝[11]では，特許情報を例に，情報検索における日本語解析の研究開発を進めていることが，報告されている．

2. 情報検索における日本語解析

情報検索システムは，

（1）　検索用ファイルを作成する情報蓄積部，

（2）　検索条件式を入力し，所望の情報を得る検索部，

（3）　シソーラスや，索引誌等の検索を援助するツールを作成する検索補助部，

に大別できる．

情報検索の運用においては，大量の情報を能率よく蓄積できることおよび，検索が柔軟に行えることが，重要である．漢字かな混り日本語文では，漢字の入力が困難であり，また分ち書きが確立しておらず単語という単位が明確でない．このため日本語文の加工を行うには，まず，単位語の決定が必要となる．この時，処理誤りを少なくしようとすると，形態素分析とか構文解析等の日本語解析が必要となる．日本語解析を必要とする主な部分は，**図-1** に示すように，

①　検索用ファイル作成における自動インデクシング，

②　検索時のシソーラス表示を意識したシソーラス作成，

③　検索補助手段としての索引誌作成，

が考えられる．②③および，自然語形式で検索条件を入力する方法も本特集中の別稿で，関連したことが述べられるので，ここでは触れない．カタカナ書き日本語文を対象にした解析[12]もあるが，以下では，漢字かな混り日本語を対象にしたものに限って話を進める．

3. 検索用ファイル作成のための日本語解析

自動インデクシングを含む検索用ファイル作成においては，日本語解析が必要とされる．本章では，

（1）　特定の関係表現の認定としての法令文の解析，

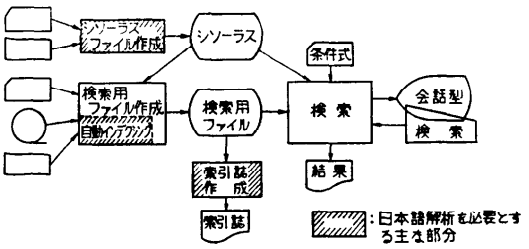

図-1　情報検索における日本語解析

† Japanese Sentence Analysis for Information Retrieval System by Hiroshi KINUKAWA (Systems Development Laboratory, Hitachi. Ltd.).

†† (株)日立製作所　システム開発研究所

（2） 汎用性を狙った論文タイトルに対するキーワード自動抽出，

（3） 日本語文構造解析による自動インデクシングの日本語解析法について，述べる．

3.1 特定形式の表現認定としての法令文の解析

法令検索は，通産省[6]や行政管理庁[10]で開発されておりその中の，関係条文検索，改廃経過検索は，法令文中の，参照法令表現や，改廃規定表現等の特定の関係を示す表現の認定の上に成り立っている．例えば改廃規定表現のうち，一部改正を示す規定文の文型は，20種程度[13]と報告され，語句も「改める」「削る」「加える」等の動詞と，「次の」「～の下・前・次に」「前段・後段」などの慣用句で表わされている．

このため慣用句テーブル照合による特定語句の認定と10種程度の述語に対する語関連表とから，改廃規定表現の認定が可能である．このように処理対象文中の特定形式の表現を認定する場合には，高度な構文分析や，意味解析によらず，小規模な語句テーブルと語関連表により，目的を果すことができる．

3.2 キーワード自動抽出のための日本語解析

情報検索システムには，検索ターム（キーワード）について，①統制語方式，②自由語方式，の二つ（両者の混合方式もある）がある．ここで，キーワード抽出を，自動化する場合には，その方式が異なってくる．①の場合には，機械辞書が必要となり，個別システム対応になりがちであるが，②の場合には，英文に対するストップ・ワード方式的な考え方により，比較的小規模なテーブルで汎用的な方式が考えられる．日本語文を対象にしたシステムでは，日本語の特性ゆえに，英文の場合ほど単純ではない．汎用性を狙ったキーワード自動抽出で実用化を目指したものとして，国文学研究資料館の国文学関係論文タイトル[14]を対象にしたシステムがある．この方式の紹介をとおして，キーワード自動抽出のための日本語解析を述べる．

（1） 単位分割：日本語文をキーワードの候補となり得る単位に分割するもので，次の2原則による．

（a） 文字種の異なるところで分割：文字種とは，

①漢字（片仮名を含む），②平仮名，③英文字，④括弧で括られた文字，⑤数字，⑥記号．

（b） 後方語テーブル，中間語テーブルと最長一致するものを分割：各テーブルには，キーワードかストップワードかの指示をしておく．

① 後方語テーブル：接尾語および共通して後尾にくる語の再分割用で，漢字列の末尾に適用．

（例） 軍記研究⟶軍記／研究

② 中間語テーブル：接頭辞，まぜ書き語，平仮名表記語を単位として認定するためのテーブル．文字列であれば，先頭，中間，末尾を問わずどの位置で一致してもよい．

（例） 我が国，

（2） キーワード抽出

（a） （1）(a)のうち①，③，④の文字列を抽出し，ストップワードファイルの語を取り除いた残りを，すべてキーワードとする．

（b） （1）(b)のうち，両テーブル中にキーワード指定のあるものを，キーワードとして抽出する．各テーブルの大きさを表-1に示す．

表-1 テーブル類の収録語数

項番	テーブル名	収録語数
1	後方語テーブル	91語
2	中間語テーブル	48語
3	ストップ・ワードファイル	382語

3.3 日本語文構造解析による自動インデクシング[15]

漢字かな混り日本語文で書かれた記事の検索システムにおいて，情報蓄積の際のキーワード付与を，現行人手作業から自動化を図り，かつ各キーワードのロール（記事文中における意味的役割，①主体，②客体，③時，④場所，⑤活動，⑥その他の主題，の6種）付けを実施し，それによって，省力化，精度の均一化を狙うものである．以下，その処理方式を説明する．

（1） 日本語文構造解析方式の概要：ロール自動付与には，文構造解析が必要であり，基本的には，名詞と述語の依存関係に着目した方式とした．ロールは次の3条件で定まる．

（a） 当該文節を支配する動詞

（b） 当該文節を構成する名詞の意味

（c） 当該文節に付く付属語

（例）「支配する」

米国が地中海を支配する．
①　④　⑤

下線：キーワード
n：ロール番号

米国が石油を支配する．
①　⑥　⑤

（2） （1）に基づき，自動インデクシングシステム

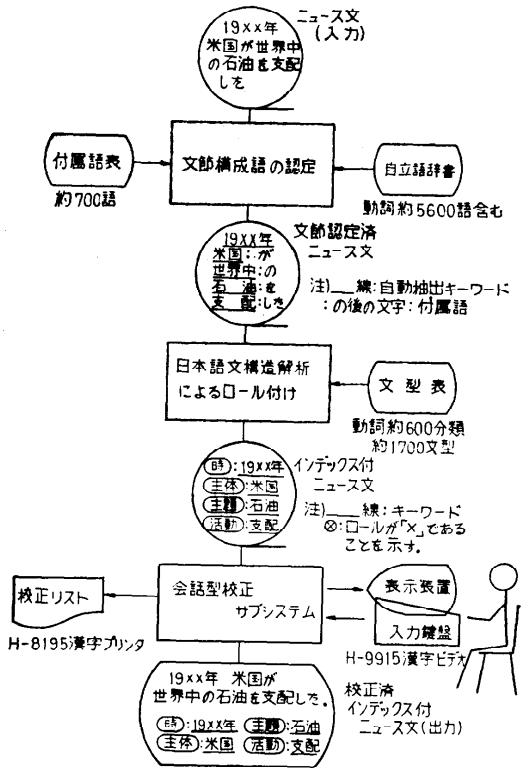

図-2 記事検索における自動インデクシング処理方式概要

を**図-2**に示す構成とした.

(a) 文節構成語の認定: 文字種の変化点で'文節'分割し, その'文節'と自立語辞書, 付属語表を照合させ, 名詞, 動詞, 付属語の認定を行う.

(b) 日本語文構造解析によるロール付与: (1)の考えに基づき作成文型表を参照して解析を行う.

(i) テーブル参照による複文構造の認定,

(ii) 受身態の能動態への変換,

(iii) 格支配関係の認定, この時, 語順が定まっていないことや語の省略に対処できる方式で, 文型表と照合,

(iv) 名詞同志, 連体形用言の修飾関係の認定,

(v) (i)～(iv)の解析結果によるロール付与,

(c) (a)(b)の自動処理結果の確認と修正のために, 漢字ビデオ端末を用いた会話型校正機能を具備した.

(3) 用語辞書, 文型表の作成に当っては, 事前に語彙調査を実施した. この結果, 収録した語数, 文型表は, 図-2に示すものとなった.

(4) プログラム容量は, HITAC-8350で, 辞書類の保守を含めて, 図-2全体で, 38K step (アセンブリ言語) 所要メモリ 132KB である.

(5) 精度は, 辞書類の出現語彙カバー率, 文型表の出現構文カバー率に大きく依存している. 実データ約1,000センテンスの実験によれば, 両カバー率を90%とすると, 文節構成語決定は, 85～90%, ロール付与率, 80～85%と推定される.

4. 考 察

(1) **3.2** の方式は, 基本的にストップワードに立脚した完全自動化を目指した方式といえる. 3種のテーブルに, 日本語の特性, システム上の制約で発生する事項 (例, まぜ書き, 平仮名書き) を吸収している.

(2) **3.3** のロール付与は, 個別システム対応のものである. このうち前半部の文節構成語の認定は, 統制語方式のキーワード自動抽出に適用可能な方式といえる.

(3) 処理テーブルに盛り込む機能が度高になれば木目の細かい処理は可能となるが, システムとしての汎用性とメンテナビリティを損なうものとなる.

(4) 自動インデクシング技術は, 自動抄録技術の基礎となるものである. 情報検索のための日本語解析にも, より高い精度を得るため処理手法の持つ精度への効果の評価を踏まえつつ, 意味解析, 談話解析の導入を, 検討していく必要があると考える.

5. おわりに

情報検索のための日本語解析に関して, 実用を目指した生データの解析を行っているものとして, 上記3事例を取り上げた. これ以外にも, より進んだ手法を開発して, 実用に供しているシステムがあるかと思うが, 特に, 検索用ファイル作成に必要とされる日本語解析の大筋は, 触れ得たと考えている. 自然語形式の問合せ方式の採用や, 推論を行う自然語質問応答システムへの接近も考えられるが, その時には, 本稿で述べた部分の他にも日本語解析が必要となる. その技術については, 本特集の別稿に述べられる部分の成果に待つものであり, 実用システムへの適用という点か

ら，評価，検討していく必要があると考える．

参 考 文 献

1) 長尾 真他2：日本語文献における重要語の自動抽出，情報処理 Vol. 17, No. 2 ('76-2).
2) 高橋達郎他2：JICST オンライン情報サービスの現状，情報管理 Vol. 20, No. 7 ('77-10).
3) (財) 日本情報処理開発協会：日本語情報処理システムの研究開発，50-S001 (昭51-3).
4) 加藤多恵子：外務省における情報検索システム：情報管理 Vol. 20, No. 10 ('78-1).
5) 塚田賢志他1：漢字 IR とそのサポートシステム，情報管理 Vol. 20, No. 4 ('77-7).
6) 電子計算機利用に関する研究会：文字情報処理システム—文書検索サブシステムの設計 (昭51-3).
7) 日本経済新聞社：記事検索システム NEEDS-IR.
8) 植村俊亮：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，747号 (昭49).
9) 国文学研究資料館：国文学研究資料館におけるコンピュータおよび漢字システム (昭53-3).
10) (株) 芙蓉情報センター：行政機関における漢字情報処理に関する調査研究報告書 (昭51-3).
11) 森 健一他3：特許請求範囲文の段落分割，情報処理学会計算言語学研究会 ('78-2).
12) 絹川博之他2：生活相談事例検索におけるキーワード自動抽出システム——国民生活センターでのケース—— 情報管理 Vol. 19, No. 2 ('76-5).
13) 佐藤，岡本：法令改正に関する日本語の処理，情報処理学会計算言語研究会 ('78-9).
14) 星野雅英：国文学関係論文タイトルからのキーワード自動抽出システムについて，第14回情報科学技術研究集会発表論文集.
15) 絹川博之他2：日本語文構造解析による自動インデクシング方式，情報処理学会プログラミング・シンポジウム予稿集 ('78-7).

(昭和54年5月30日受付)