U.D.C. 681.322.056:002.54:809.56-15

日本語情報検索システムにおけるキーワード自動抽出

Automatic Indexing Methods for Information Retrieval System of Japanese Text

絹川博之* *Hiroshi Kinukawa*
田中和明* *Kazuaki Tanaka*
池上信男** *Nobuo Ikegami*

日本語文献の検索システムで，キーワード付与作業の省力化と品質の均一化を目的とした自動インデクシング方式として，日本語文構造解析方式と不要語除去方式を設定し，プロトタイプによる実験を行なった。前者は，キーワードの用語を限定した検索方式での利用を前提としており，各用語の文法・意味情報を用いた日本語文の構造解析により，キーワードとその文中での位置づけや意味をも合わせて抽出する方式で，抽出精度は80～85%であった。後者は，キーワードの用語を限定しない検索方式での利用を前提にしており，汎用性，高速性，保守性及びキーワードの抽出漏れの防止を主目的に考えられた方式で，抽出精度は約95%であった。

1 緒　　言

日本語ワードプロセッサを含む漢字入出力機器の普及と日本語情報処理技術の進歩により，漢字や平仮名を含む日本語情報，文書情報を対象にした情報検索システム建設のニーズが高まっている。実際，日本科学技術情報センターでは，JOIS-II(JICST Online Information System II：科学技術情報のオンライン情報検索システム)[1]を，日本特許情報センターでは，PATOLIS(Patent Online Information System：特許情報検索システム)[2]を，日本経済新聞社では，NEEDS-IR(Nikkei Economic Electronic Data-Bank Service：新聞記事検索システム)[3]をそれぞれ開発し，サービス業務を開始している。また，企業内では，これらの社外サービスとの連動及び社内OA(Office Automation)の一環として，特許情報，設計情報及び文書情報の検索システムの構築が各所で進められている。

このような状況を踏まえて，日立製作所では，汎用日本語情報検索ソフトウェアとして，HITAC Mシリーズ計算機下で動くORION(Online Retriever of Information)を開発し製品提供を開始しているが，機能の充実が望まれている。例えば，単に漢字をサポートしているだけでは十分でなく，使いやすく効率よく運用できる機能の具備が求められている。具体的には，情報検索では大量情報の蓄積が前提であることから，情報入力蓄積作業の省力化が必要である。この作業のうちインデックス付与作業は，入力情報を解釈加工する作業であり，単なる情報入力作業とは質の違った作業である。このため，情報入力作業の省力化という点から，このインデックス付与作業の自動化がまず考えられ，このためのツール(自動インデクシング)が求められている。また，情報検索のインデックス(例えば，キーワード)は，大量情報の中から所望情報を短時間で探し出すための媒体であり，このインデックス付与精度が検索精度を大きく左右している。しかし，従来ではインデックス付与作業を人手に委ねていたため，付与作業者の個人差があってインデックス付与の品質の確保が難しいものとなっていた。このため，インデックス付与の個人差の解消と品質の確保という点からも，インデックス付与自動化ツールが求められている。更に，OAの進展に伴いオフィス各所で発生する電子化された文書の検索を可能とすることが要請されている。このためには，検索用のインデックス(例えば，キーワード)の付与作業の自動化ツールを構成機能として，文書検索システムに組み込むことが必要となってきている。日立製作所では，これらの要請に応ずるため，日本語文構造解析方式と不要語除去方式の自動インデクシングの研究開発を進めている。

以下，プロトタイプによる実験を踏まえて，方式の特徴，適用対象などについて述べる。

2 日本語情報検索システムにおける位置づけ

自動インデクシングシステムは，キーワード自動抽出処理と，その校正処理を行なう各サブシステムから構成され，その方式設定に当たっては，検索系と蓄積系から成る検索システム全体にわたった連係を十分配慮する必要がある。

図1に，日本語情報処理システムの全体像を示す。

キーワード検索方式としては，次の2タイプ[4]がある。

(1)　標準ターム方式

検索対象は限定されるが，キーワードとなる用語を標準化して，表記の違いや同義語による検索再現率※1)(検索漏れ率の補数)の低下を防ぐ方式で，用語の文法・意味情報をもとに，キーワードの文中での位置づけや意味をもインデックスとして合わせて検索することで，検索適合率※2)(検索ノイズの補数)の向上を図ることも可能である。

(2)　フリーターム方式

キーワードとなる用語を制限しない方式で，処理方式が適用対象に依存せず，新語の取込みも容易である。ホストコンピュータの演算処理やファイルアクセス処理の高速化に伴い，検索質問中の検索キーと同義語であるキーワードをもつすべての文献を検索できる機能の実用化により，最近見直されてきている。

これらの検索方式に対応する自動インデクシング方式として，次の2方式を設定した。

※1)　検索再現率：$\frac{\text{(検索文献中の適切な文献数)}}{\text{(文献中で，本来検索されるべき文献数)}}$

※2)　検索適合率：$\frac{\text{(検索文献中の適切な文献数)}}{\text{(検索された全文献数)}}$

* 日立製作所システム開発研究所　** 日立製作所ソフトウェア工場

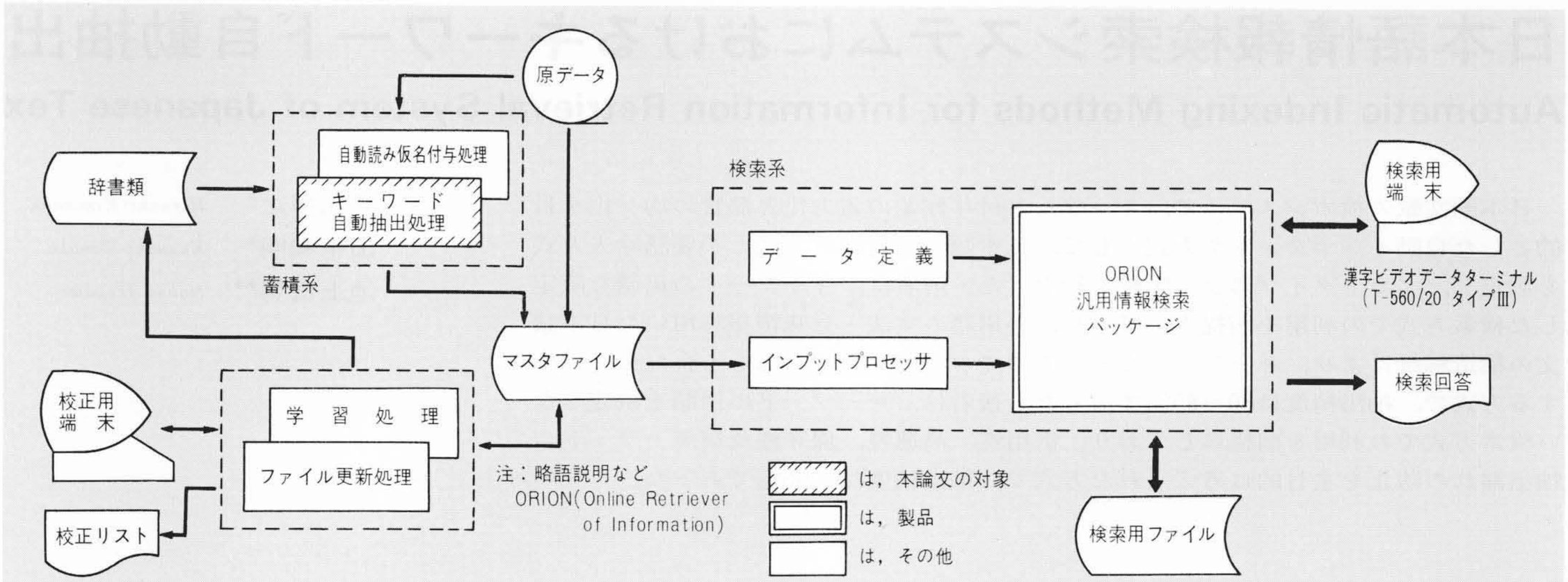

図1 日本語情報検索システムの全体像 日本語情報検索システムは，ORIONを核とする検索系と，自動インデクシング処理やファイル更新処理を含む蓄積系から構成される。

(1) 日本語文構造解析方式

あらかじめ登録されたキーワードとなる言葉と，その文法・意味情報をもとに，日本語文の構造解析を行なってキーワードを選定する方式である。

(2) 不要語除去方式

汎用性，高速性，保守性，検索漏れの防止を主目的に考え，不要語が多少含まれていても，キーワードとなる言葉を漏れなく抽出する方式である。

次章では，この両方式について述べる。

3 自動インデクシング方式

3.1 日本語文構造解析方式による自動インデクシング[5)]

漢字仮名まじり日本語文からのキーワード抽出と，各キーワードのロール(文中での意味的役割：① 主体，② 客体，③ 時，④ 場所，⑤ 活動，⑥ その他の主題，の6種)付与を自動化し，それによって，省力化と検索適合率の向上をねらうものである。以下，その処理方式について説明する。

(1) 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，付属語の認定を行なう。

(2) 日本語文構造解析によるロール付与

(a) 日本語文型表を参照して，文構造を認定する。

(i) 複文パターン表参照による複文構造の認定

(ii) 受身態から能動態への変換及び格の標準化

(iii) 日本語文型表参照による格支配関係の認定

(iv) 名詞同士，連体形用言の修飾関係の認定

(b) 次の条件により，ロールを付与する。

(i) 当該文節の支配する動詞

(ii) 当該文節を構成する名詞の意味

(iii) 当該文節に付く付属語

〔例〕「支配する」

米国(①)が地中海(④)を支配(⑤)する。

米国(①)が石油(⑥)を支配(⑤)する。

{下線：キーワード，(n)：ロール番号

(3) 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て対話形で校正を行なう。

当方式の処理手順を**図2**に示す。

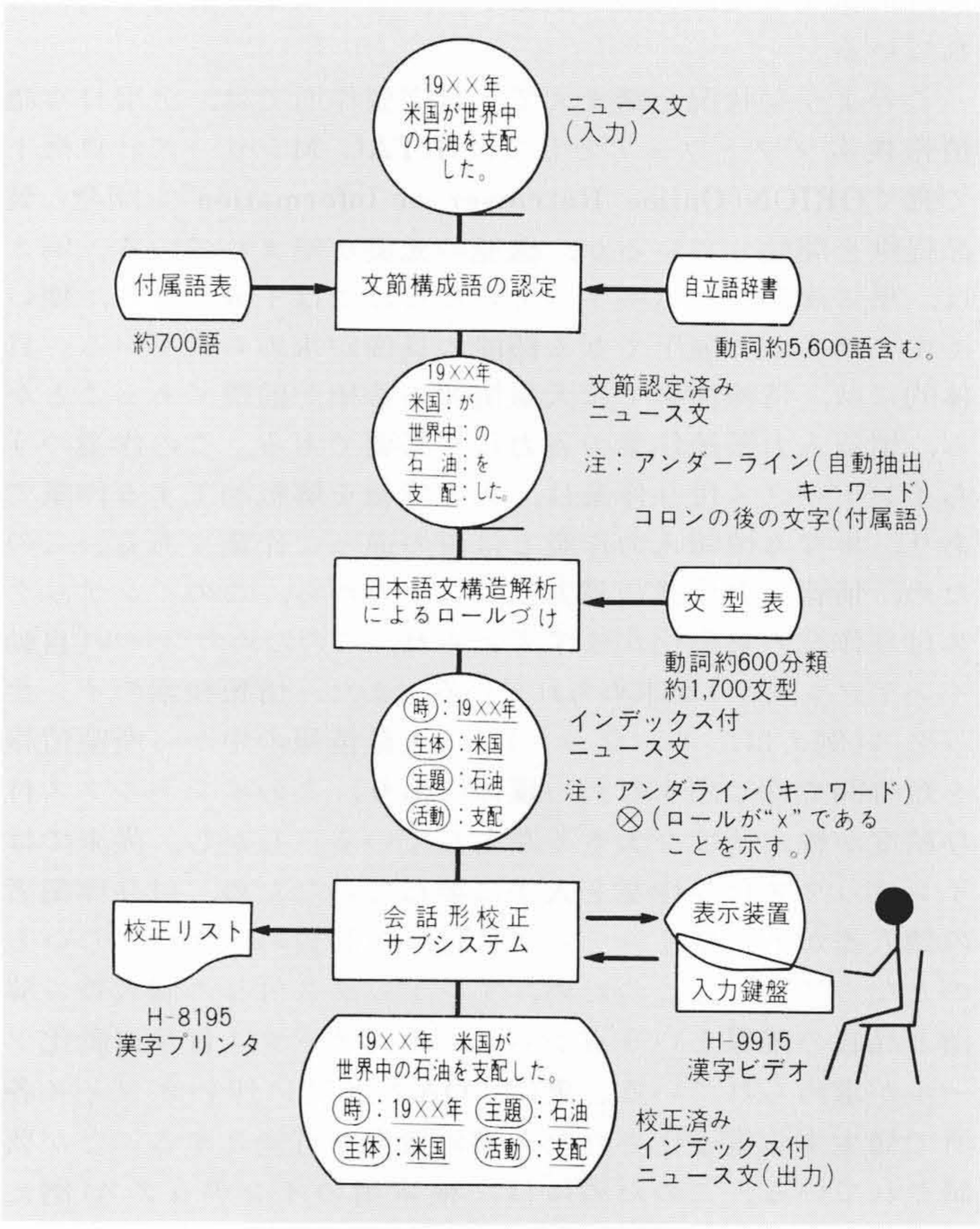

図2 日本語文構造解析方式による自動インデクシング この方式の処理手順は，次の三つに大別される。(1) 文節構成語の認定，(2) 日本語文構造解析によるロールづけ，(3) 会話形式による自動抽出処理結果の校正

3.2 不要語除去方式による自動インデクシング方式

英文情報を対象とした情報検索システムでは，英文中から不要語を除去し，残りの単語をキーワードとする方式が実用化されているが，日本語文では，単語単位で分かち書きされておらず，また，明確な分割規則がないことから，独自の方式を検討する必要があった。以下，処理方式について説明する。

(1) 漢字仮名まじり日本語文の自動分かち書き

日本語文で重要な言葉は，非平仮名で書き表わされているという日本語の特性を利用して，文字列の分割を行なう。

(2)　自立語の認定

付属語表を用いて，分かち書きされた文節から付属語を除去することにより自立語を認定する。

(3)　不要語の除去

不要語テーブルの登録語や，キーワードの文字列構成基準を満足しない自立語を除去する。

処理手順を**図3**に示す。

ここでは，キーワードを，**図4**の斜線で囲まれた領域内の文字列と定義している。

4　プロトタイプによる実験・評価

前**3**章で述べた両方式の実現性を検討するため，**表1**に示す文献を対象としたプロトタイプを作成し，実験・評価を行なった。

各プログラムはアセンブリ言語で記述されており，その構成，ステップ数，所要容量及び1センテンス当たりの平均処理時間は，**表2**に示すとおりである。

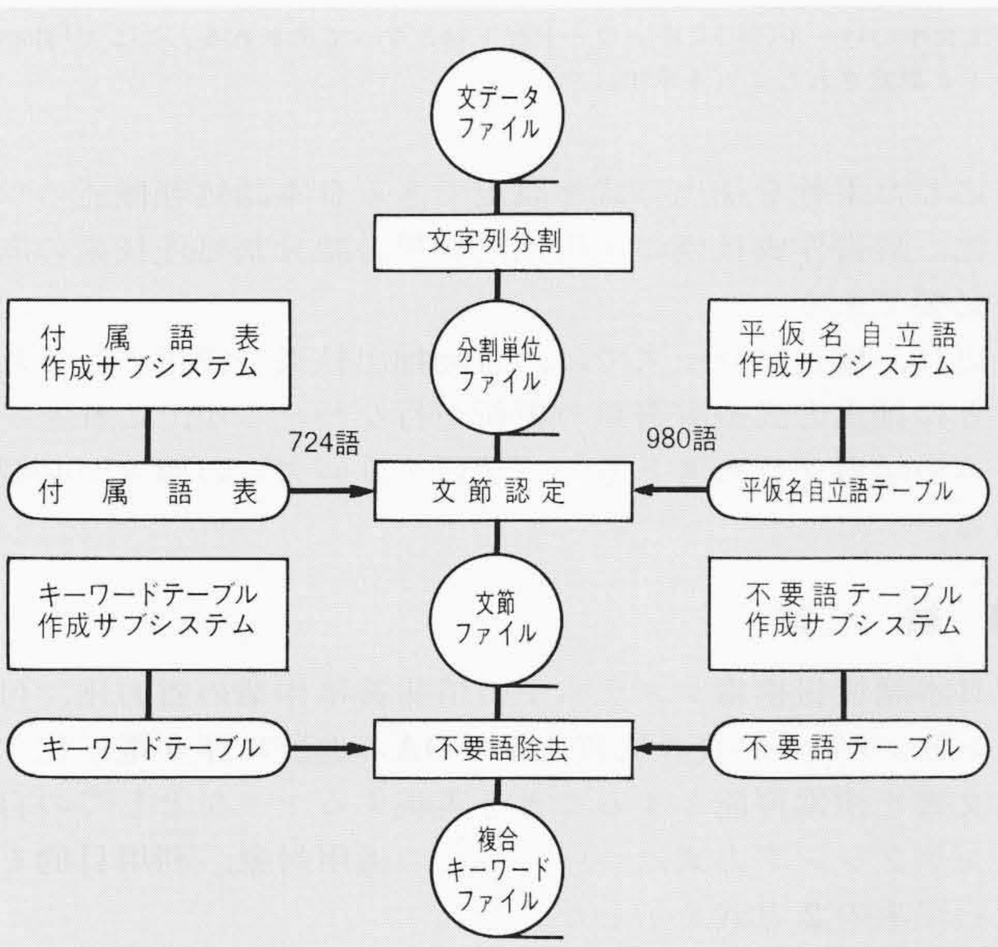

図3　不要語除去方式による自動インデクシング　この方式の処理手順は，次の三つに大別される。(1)漢字仮名まじり日本語文の自動分かち書き (2)文節の中から自立語を認定する。(3)不要語の除去

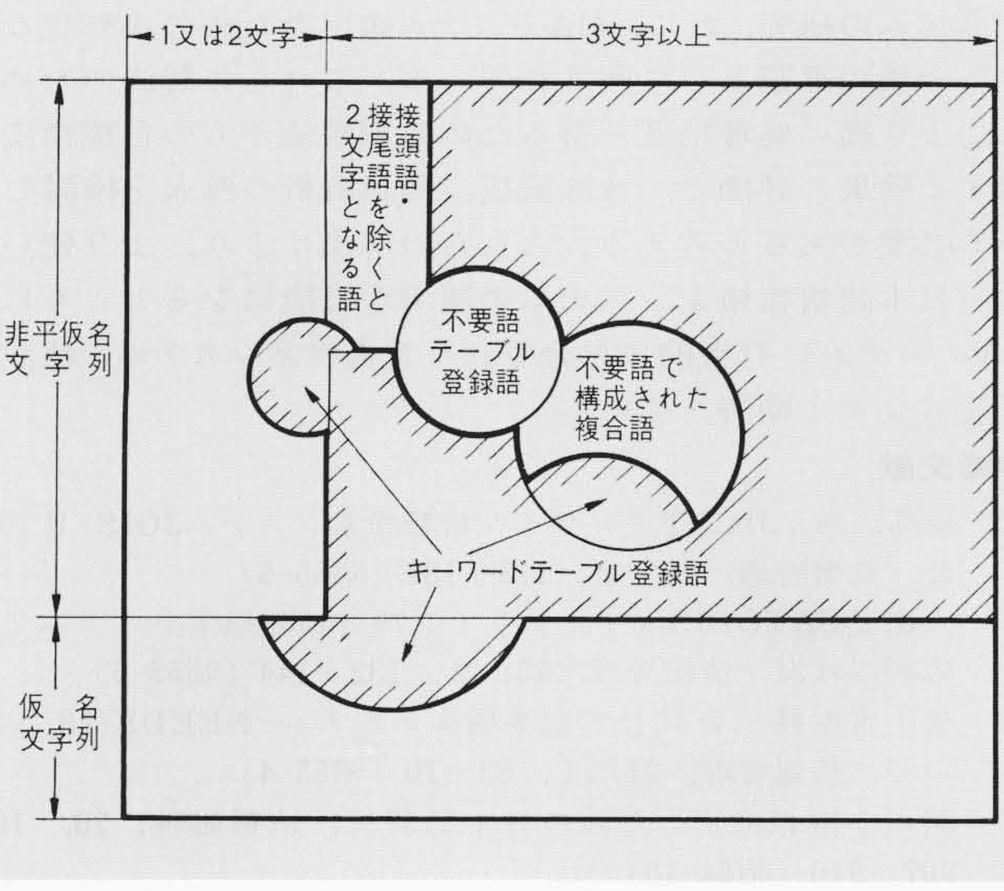

図4　不要語除去方式におけるキーワードの定義例　斜線で囲まれた部分をキーワードとする。

表1　自動インデクシングの適用データ　自動インデクシング方式の適用データとして，日本語文構造解析方式については外電記事文を，不要語除去方式については「日立評論」のタイトルと要旨を用いた。

自動インデクシング方式	実験対象データ	報告書数	センテンス数
日本語文構文解析方式	外電記事文	281	1,225
不要語除去方式	「日立評論」のタイトルと要旨	78	361

表2　自動インデクシングのプロトタイプの構成と性能　各処理プログラムは，アセンブリ言語で記述されている。処理時間はHITAC M-180換算で，日本語文構造解析方式が合計で435ms/センテンス，不要語除去方式が合計で65.1ms/センテンスである。

方式	処理プログラム名	ステップ数	所要メモリ	性能	備考
日本語文構文解析方式	文節構成語の認定処理	3kS	60kバイト	2,640ms/センテンス	使用計算機 HITAC 8350
	文構造解析によるロール自動付与処理	12kS	120kバイト	650ms/センテンス	
不要語除去方式	文字列分割処理	1.9kS	13.3kバイト	8.8ms/センテンス	使用計算機 HITAC M-180
	文節認定処理	1.4kS	40.3kバイト	14.4ms/センテンス	
	不要語除去処理	1.1kS	7.3kバイト	41.9ms/センテンス	

注：略語説明　kS(キロステップ)

500件/日(10万件/年)のデータ更新を考えると，自動インデクシング処理時間はHITAC M-180換算で，日本語文構造解析方式ではCPU(中央処理装置)処理時間で約15.8分，実行時間で約1時間，不要語除去方式では，それぞれ約2.5分，約10分程度であり，大量のデータ更新には後者が適している(なお，各処理時間は，原理実験を目的としたプロトタイプによるものである)。

図5に，抽出結果の編集出力様式を示す。

表3，**4**に，抽出処理精度を示す。

日本語文構造解析方式における精度は，辞書類(各用語に関する品詞情報や意味分類情報などが登録されている。)の出現語いカバー率※3)，文型表の出現構文カバー率に大きく依存しており，事前の語い調査によって収録された用語辞書や文型表(**図2**参照)をもとにした外電記事文データ，1,225センテンスの実験では，両カバー率90%，文節構成語決定精度※4) 85～90%，ロール付与精度80～85%程度であった。

本方式は抽出されるキーワードに，主体，客体といった文中での位置づけや，場所，時などの意味情報が付加されるので的確な検索に役立つが，一般には，日本語文の表現形態の多様さから，辞書や文型表の追加・整備に追われることが予想され，分野限定形の利用方法となる。

不要語除去方式の「日立評論」タイトルと要旨，361センテンスに対する適用実験では，文字列分割精度98.5%，文節認定精度95.3%程度であった。本方式の実用化には，画一的なキーワードの定義に従わないキーワードや，不要語に対する例外辞書の充実及び正確な複合語分割技術が必要となる。今

※3)　カバー率：$\frac{(\text{出現語中の辞書表収録延べ数})}{(\text{出現語延べ数})}\times 100$

※4)　精　　度：$\frac{(\text{正処理数})}{(\text{対象総数})}\times 100$

記事番号： １１１０２２５○ ア　文番号： 1
＜入力原文＞
７７０１０７カーター次期大統領，景気対策大綱を発表

＜処理＞　ＩＤ作成区分（ ）

ＬＫ	番号	Ｒ	キーワード	番号	Ｒ	キーワード	番号	Ｒ	キーワード
主文	1	①	カーター	2	①	次期大統領	3	⑥	景気
	4	⑥	対策	5	⑥	大綱	6	⑤	発表
非パ	7	③	７７０１０７						

(a) 日本語文構造解析方式の例

プロセス制御システムは複雑化，高級化する傾向にあり，その設計，エンジニヤリング，計装工事などに多大の時間を必要としている。この解決のために，アナログ制御，直接ディジタル制御，シーケンス制御及び計算機制御を合理的に結合した総合計装システム，「ユニトロールシグマシリーズ」を開発した。「ユニトロールシグマシリーズ」は，従来の電子式工業計器「ユニトロールＥシリーズ」を総合計装システムの立場から見なおしてモジュール化するとともに，マイクロコンピュータを集合計器の形で導入し，性能の拡大，結合の合理化を図ったものである。本稿は，その全体の構成，方式について紹介する。

(b) 不要語除去方式の例

図５　自動インデクシング適用例　日本語文構造解析方式の適用例では，同一主文中のロール(ⓝ)とキーワードが１対となって示される。ここで「非パ」とは，文型表によらないロール付与不要語除去方式では，アンダーライン部分がキーワードと認定されたことを示す。

表３　日本語文構造解析方式による自動インデクシングプロトタイプの処理精度　表１の外電記事文に対して，出現語いカバー率，出現構文カバー率がそれぞれ90%であるときの処理精度である。

項番	評価項目		精度(%)
1	文節構成語認定		85～90
2	ロール付与	複文構造認定	94～98
3		格支配関係認定	86～91
4		体言支配関係認定	90～95
5		ロール付与精度	80～85

表４　不要語除去方式による自動インデクシングプロトタイプの処理精度　表１の「日立評論」のタイトルや要旨に対する処理精度である。

項番	評価項目		精度，語数
1	文字列分割精度		98.5%
2	文節認定精度		95.3%
3	キーワード	キーワードと正しく認定されたもの	4,384語
4		不要語と誤認されたもの	136語
5		小計	4,520語
6	不要語	不要語と正しく認定されたもの	9,470語
7		キーワードと誤認されたもの	3,573語
8		小計	13,043語
9	キーワード認定精度	(項３)/(項３)+(項４)	97.0%
10	キーワード認定適合率	(項３)/(項３)+(項７)	55.1%
11	不要語除去精度	(項６)/(項６)+(項７)	72.6%
12	キーワードテーブルに収録する必要のある語数		51語
13	不要語テーブルに収録する必要のある語数		1,972語

回の実験では，例外辞書を用意しない場合，キーワードとほぼ同数の不要語がキーワードと誤認された。しかし，本方式は，適用分野によって例外辞書の登録語は異なるが，抽出アルゴリズムに分野依存性がなく，また人による校正情報をもとにした辞書の自動登録が可能であり，汎用性，保守性に富むシステムに拡張できると考えられる。

プロトタイプは，特定分野の日本語文を対象とする専用システムとして試作され，自動インデクシング方式の実現性を確認することができたが，実用化を図るためには，適用対象に応じた柔軟な抽出方式を設定できる日本語処理機能のツール化，辞書作成技術の汎用化及び複合語分割処理技術の向上が必要である。

更に，校正プロセスでは，自動抽出結果の校正と，それをもとに抽出方式や辞書類の更新が行なわれるが，これをシステムの学習プロセスと考え，学習メカニズムの導入が必要である。

5 結　言

日本語情報検索システムでの情報蓄積作業の省力化，付与インデックスの品質の均質化及びOAの進展に伴う電子化された文書を検索可能とすることを実現するツールとしての自動インデクシング方式について，その適用対象，利用目的を考慮して次の２方式を示した。

(1) 日本語文構造解析方式
(2) 不要語除去方式

更に，両方式についてプロトタイプによる実験を行ない，それらの実現性を確認した。両方式の実用化には，自動インデクシングのための複合語分割，同義語認定などの日本語処理技術のツール化，抽出処理及び辞書の拡充のための学習メカニズムの研究，更に，OAシステム適用のための小形化などが，今後の課題として残されている。これらの解決のためには，より高い処理精度を得るため処理手法のもつ目標精度に対する効果の評価と，意味解析，談話解析の導入を検討してゆく必要があると考える。これらの解決により，より使いやすい日本語情報検索システムの運用が可能になるとともに，OAシステムに有機的に結合された文書検索システムの建設が可能になると期待している。

参考文献

1) 福島，外：JICSTオンライン情報検索システムJOIS-IIの概要，情報管理，**23**，２，118～131（昭55-5）
2) 大山：PATOLIS漢字オンライン特許情報検索システムとその利用状況，情報管理，**23**，２，132～144（昭55-5）
3) 菅：日本経済新聞社の記事検索システム「NEEDS-IR」について，情報管理，**21**，１，63～70（昭53-4）
4) 絹川：情報検索のための日本語解析，情報処理，**20**，10，907～910（昭54-10）
5) 絹川，外：日本語文構造解析による自動インデクシング方式，情報処理学会論文誌，**21**，３，200～207（昭56-5）