Vol. 27 No. 1 情報処理学会論文誌 Jan. 1986

日本語科学技術文における専門用語の自動抽出システム†

吉村賢治†† 日高達††† 吉田将†††

科学技術文献の機械翻訳に関する研究の進展に伴い，これに用いる専門用語辞書の開発が望まれている．しかし，専門用語はその数が膨大である上に新しい用語の出現数も多いことから，人手だけでこれを収集することは困難であり，収集の機械化が必要である．本論文では，日本語の専門用語を自動収集するために作成した，日本語科学技術文から専門用語を自動抽出するシステムについて報告している．この専門用語の自動抽出システムでは，まず付属語辞書と一部の自立語を収めた小規模な特殊自立語辞書を用いて粗い形態素解析を行ったのち，文法情報を利用して専門用語の候補を求め，次に一般用語辞書を参照して専門用語を抽出するという方法を用いている．この方法の有効性については，日本科学技術情報センターの科学技術文献速報を対象に実験を行い，入力文中の専門用語の98.3%を，97.1%の精度で抽出できることを確認した．

1. まえがき

科学技術文献の機械翻訳や機械支援翻訳に関する研究の進展に伴い，これに用いる専門用語辞書の開発が望まれている．しかし，専門用語は，その数が膨大である上に新しい用語の出現数も多いことから，人手だけで収集することは困難であり，専門用語収集の機械化が必要である．本論文では，科学技術文献の日本語表題と抄録文から専門用語を抽出するシステムとその実験結果について報告する．

日本語文からの用語抽出については，国立国語研究所の語彙調査に関する先駆的研究[1)-4)]があるが，人手による前処理を必要としている．用語を自動抽出する際の技術的な問題は，(Ⅰ)入力文を適切な単位に分割する，(Ⅱ)各単位の必要性を判定する，の2点である．これまで，植村の自動索引システム[5),6)]は字種情報だけで二つの処理を行っており，長尾らの重要語の自動抽出システム[7)]では，(Ⅰ)を字種情報を用いて行い，(Ⅱ)を χ^2 検定法を応用した統計処理で行っている．しかし，科学技術文には仮名書きの用語も多数あり，字種情報だけを用いた処理には精度上の問題がある．また，必要性の判定に χ^2 検定法を応用した場合，複数分野の文献を同時に処理する必要がある．一方，荒木らの論文表題からキーワードを自動抽出するシステム[8)]では，不要語列のテーブルを用いて(Ⅰ)，(Ⅱ)の処理を同時に行い，97.5%の精度を得ており，科学技術文から専門用語を抽出する場合にも，入力文から不要語を除去する方法が有効であると考えられる．この場合，どれだけの不要語列を用意するかで抽出の精度が決まるが，不要語となりうる自立語の数からも推定されるように，一般の科学技術文における不要語列の数は膨大となり，テーブルを用意することは困難である．そこで，本論文では，一般用語だけからなる国語辞書を用いて形態素解析を行い，未登録語となった専門用語を抽出するという方法を提案している．この方法では，不要語列を構成する要素の辞書と要素間の連接規則を用いて，不要語列の集合をより網羅的に規定できる．

本システムでは，文献9）で報告した未登録語を含む日本語文の解析が可能な形態素解析アルゴリズムを用いており，この特徴を利用して二次記憶上にある自立語辞書の検索に要する時間の削減を図っている．なお，ここで抽出の対象としているのは名詞の専門用語だけである．

2. システムの概略

専門用語を自動抽出する過程を**図1**に示す．最初に主記憶上に常駐可能な特殊自立語辞書と付属語辞書を用いて入力文の形態素解析を行う．この結果，一般の自立語は未登録語となる．次に名詞の可能性をもつ未登録語をキーとするKWICレコードを作成する．その後，このKWICレコードの各キーについて接辞の除去，一般用語の除去および切断誤りの回復と読み振りを行って専門用語の候補を得る．以下，図1の各ステップについて述べる．

† An Automatic Extraction System of Technical Terms from Japanese Scientific Documents by KENJI YOSHIMURA (Department of Electronics, Faculty of Engineering, Fukuoka University), TORU HITAKA and SHO YOSHIDA (Department of Electronics, Faculty of Engineering, Kyushu University).

†† 福岡大学工学部電子工学科
††† 九州大学工学部電子工学科

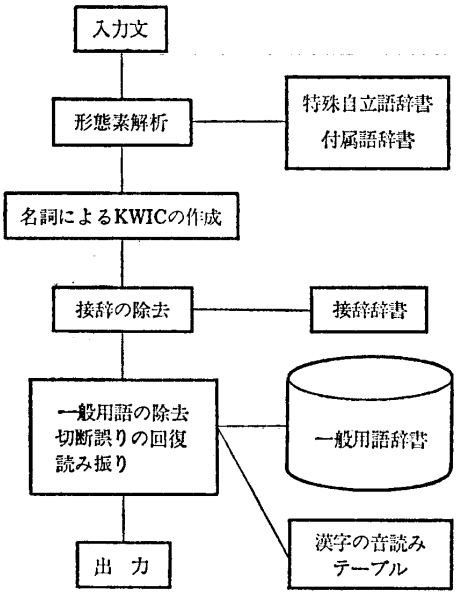

図 1 専門用語自動抽出の過程
Fig. 1 Process of automatic extraction of technical terms.

3. 形態素解析

3.1 文法モデル

形態素解析で用いている文法モデルを BNF 記法を使って示す．ここで，“〔” と “〕” で囲んだものが解析の単位であり，それ以上分解されないものである．これを解析単位と呼ぶ．

〈文〉::=〈文節〉|〈文〉〈文節〉
〈文節〉::=〈自立部〉|〈自立部〉〔活用語尾〕
|〈文節〉〔付属語〕
〈自立部〉::=〔正書表記の自立語〕|〔数詞〕
|〔非正書表記の自立語〕|〔漢字〕
|〔平仮名〕|〔記号〕|〔連結記号〕
|〔アルファベット・片仮名列〕

この文法モデルでは，文を単なる文節の並びとしてとらえ，文節が文を構成するための条件は規定していない．ただし文節を構成する解析単位の並びは，連続する二つの解析単位間の接続条件で規定する．自立部のうち，〔正書表記の自立語〕，〔非正書表記の自立語〕，〔数詞〕は単語であるが，〔漢字〕，〔平仮名〕はおのおの一文字の漢字と平仮名を意味しており，入力文に未登録語が存在した場合に，これを構成する文字をそれぞれ一つの自立語として解析を行うために用意している．〔アルファベット・片仮名列〕も同様に外来語の未登録語に対処するための解析単位であり，一連の英字列または片仮名列を意味している．〔記号〕は句読点など自立語の一部とならない記号で，〔連結記号〕はハイフンなど複合語の形成に使用される記号である．この文法モデルでは，これらの記号類も一つの文節として扱う．

3.2 形態素解析の手順

形態素解析では，参考文献 9)，10) で報告した表方式のアルゴリズムを用いている．このアルゴリズムでは，最初に入力文を文頭から文末に向かって走査し，入力文を構成しうるすべての解析単位を要素とする表を作成し，次にこの表を参照して入力文の分割を決定する．このとき，3.1 節で述べた文法モデルは正しい文の構造を規定する能力が弱いため，作成した表からは誤ったものも含む多くのあいまいな分割が作り出される．このシステムでは，それらの分割からもっともらしいものを取り出すために次のヒューリスティクスを用いている．

(H1) 文節数が最小となる分割を優先する[10]．
(H2) 解析単位の個数が最小となる分割を優先する．
(H3) 非正書表記の自立語の個数が少ない分割を優先する．
(H4) 未登録語の総文字数が少ない分割を優先する．

ヒューリスティクスの基本的な強さは (H1)，(H2)，(H3)，(H4) の順とする．これらのヒューリスティクスを一つの尺度で評価して処理を簡潔にするため，解析単位に種類に応じたコストを設定する．このコストは，入力文を構成する解析単位のコストの総和が最小のときヒューリスティクスを最も満足するように設定する．各解析単位のコストを**表 1** のように表すと，(H1)，(H2) に従うために，

$C_M, C_K, C_R, C_H, C_J, C_N > C_G, C_F > 0$

表 1 解析単位の種類
Table 1 The variety of units of analysis.

解析単位	記号	コスト
i 正書表記の自立語	J	C_J
ii 数詞	N	C_N
iii 付属語	F	C_F
iv 活用語尾	G	C_G
v 非正書表記の自立語	H	C_H
vi 漢字	K	C_K
vii アルファベット，片仮名列	R	C_R
viii 平仮名	M	C_M
ix 連結記号	C	C_C
x 記号	S	C_S

とする．次に（H3）に従うために，

$C_H > C_J, C_N$

とし，（H4）に従うために，

$C_M, C_K, C_R > C_J, C_N$

$C_M > C_H$

とする．また，このシステムでは4章で示すように名詞の可能性がある文字列を抽出することが目的であるため，

$C_G > C_F$

とする．以上のことと，字種の違いにより同一の文字列がなりえない解析単位の種類等を考慮して，各解析単位のコストは次の関係を満たすように設定する．

$C_M > C_H, C_K, C_R > C_J, C_N > C_G > C_F > 0$

ここで，〔記号〕および〔連結記号〕は他のいかなる解析単位とも競合しないため，C_S, C_C の値は任意である．

なお，本論文で報告する実験は次に示す値で行った．

$C_C = C_S = 0$, $C_F = 1$, $C_G = 2$, $C_J = C_N = 4$,

$C_H = C_K = C_R = 6$, $C_M = 14$.

これらの値は，上記の関係だけを満たすように設定したものであり，値自体に理論的根拠はない．ただし，C_M は形態素解析の実験から，$2C_H < C_M + C_F < 3C_H$ となるように設定している．

参考文献9)，10)で示したアルゴリズムでは，作成した表から入力文の解析結果を抽出するが，本システムでは，4章で述べるように表から名詞（列）をキーとした KWIC レコードを作成する．

3.3 特殊自立語辞書

参考文献9)の形態素解析システムでは単語数約83,000，見出し語数 170,000 の自立語辞書[11]を使用しており，処理時間の約半分を二次記憶上にある自立語辞書のアクセスに費やしている．そこで，本システムにおける形態素解析では，上記の辞書の代りに主記憶上に常駐可能な小規模な自立語辞書を使用して，処理時間の削減を図っている．この辞書を特殊自立語辞書と呼ぶ．特殊自立語辞書には，専門用語やその構成要素になりえない連体詞，接続詞，副詞，形容詞などの814単語が登録されている．その内訳を**表2**に示す．なお，特殊自立語辞書の作成にあたっては首藤のデータ[12]を参考にしており，多くの慣用的表現を単語として登録している．

この特殊自立語辞書中の単語は，切断禁止文字列を除いて不要語であり，形態素解析の結果，専門用語を含む一般の自立語は未登録語となる．ここで切断禁止文字列とは，副詞“絶対”によって切断されては不都合である“絶対値”のような文字列であり，解析単位の種類は漢字として扱う．

表2 特殊自立語辞書の内訳
Table 2 Details of exceptional independent words dictionary.

品　詞	語数
副詞	335
連体詞	209
接続詞	160
名詞	67
形容詞	35
切断禁止文字列	8

4. 名詞をキーとした KWIC の作成

形態素解析で作成した表を参照し，コストの総和が最小になる分割を文末側から文頭に向かって走査し，名詞の可能性がある未登録語をキーとする KWIC レコードを作成する．キーになる解析単位の列をXで表すと，Xは BNF 記法を用いて次のように定義される．ここで G* は五段動詞の連用形活用語尾を表しており，その他の記号は表1に示した解析単位に対応している．

X::=XX

X::=XCX

X::=NCX

X::=NX

X::=XG*

X::=K|M|R

この定義を満たす最長のXを KWIC レコードのキーとする．なお，この段階で一文字の平仮名がキーになっている KWIC レコードは削除している．

この段階で抽出したものは，特殊自立語辞書に登録している名詞以外の名詞または名詞の列である．ここでは，JICST（日本科学技術情報センター）発行の科学技術文献速報・電気工学編 Vol. 25, No. 6 の先頭から取り出した100文献の表題および抄録文について行った名詞抽出の実験結果を示す．

対象：JICST・科学技術文献速報
　　電気工学編，Vol. 25, No. 6
　　文献番号　E82060001～E82060100

総文数	428 個
名詞の総数 (a)	2339 個
抽出できなかった名詞の個数 (b)	66 個
誤って抽出した文字列の個数 (c)	124 個

磁気しゃ へい (1)
二乗平均平方根的広 がり (2)
じょう 乱解析 (3)
不変はめ 込み理論 (4)

図 2 抽出に失敗した名詞の例
Fig. 2 Examples of noun that could not be extracted.

ここで,

抽出率$=(a-b)/a$

抽出の精度$=(a-b)/(a-b+c)$

と定義すると，この段階における名詞の抽出率は 97.2 %，抽出の精度は 94.8% である．抽出に失敗した名詞はほとんどのものが**図 2** に示すような漢字と平仮名で混ぜ書きされたものである．ここで，下線部がキーを示している．また，誤って抽出した文字列の大半は形容動詞の語幹であった．

5. 接辞の除去

3章で述べた形態素解析では接辞処理を行っていないため，不要な接辞が連接した名詞はそのままの形でKWICレコードのキーになっている．この形のままで次章に述べる一般用語の除去を行った場合，接辞の連接した一般用語を除去することができない．また，専門用語に接辞が連接したものも専門用語の候補となる．そこで，一般用語の除去を行う前に，接辞辞書を用いてキーの先頭にある不要な接頭語を前の文脈文字列へ，キーの末尾にある不要な接尾語を後ろの文脈文字列へおのおの移動する．このとき，名詞の一部が接辞と一致してキーから取り除かれることもあるが，この誤りのほとんどは6章のステップで回復できる．**表 3** に実験で用いた接辞辞書の内容を示す．これらの接辞は，科学技術論文速報の 10,000 文から収集した接辞のうち‘性’のように専門用語の一部となりうるものを除いた残りである．また，“以後”，“以降” などのように特殊自立語辞書にある単語も接辞辞書から除外している．なお，表 3 に示したように接辞の数が少ないが，文字列の照合だけで接辞を除去している本システムの場合，接辞数の増加がそのまま精度の向上に結びつくとは限らない．例えば，専門用語から誤って接辞を除去したために，その用語がまったく抽出できなくなる可能性もあり，接辞辞書の作成に関しては他の語との関連などに対する十分な配慮が必要である．4章で述べた名詞の抽出実験の出力に対して接辞の除去を行った場合，一般用語と専門用語について**表 4** に示す個数の接辞が除去された．接辞除去の効果については，6章で専門用語の抽出率，抽出の精度と関連して述べる．なお，誤って除去した接辞はすべて切断誤り回復のステップで回復されている．

表 3 接辞辞書
Table 3 The dictionary of affix.

接頭語	接尾語	
各	やすい	全体
諸	下	中
当該	以外	等
本	間	内
両	時	後半
	上	分
	側	用

表 4 接辞除去の回数
Table 4 Number of affix removing actions.

	一般用語から	専門用語から
正しい除去	24	53
誤った除去	34	29

6. 一般用語の除去と切断誤りの回復

6.1 一般用語辞書

このステップまでに作成された KWIC レコードのなかには，一般用語がキーになっているものもある．ここでは，一般用語辞書を用いて，このようなKWICレコードの削除を行う．また，形態素解析，接辞除去の段階で生じた専門用語の切断誤りの回復も同時に行っている．

一般用語辞書は，参考文献11)の自立語辞書から抜き出した正書表記の名詞だけで構成した．この一般用語辞書は，見出し語数が79,102個で，文頭側からの検索に用いる正引き辞書と文末側からの検索に用いる逆引き辞書からなる．辞書のデータ構造は拡張B-tree[13]と TRIE[14]を併用した構造である．容量は，正引き辞書，逆引き辞書の見出しファイルがともに1,410kbyte，両辞書共用の読みファイルが4,371kbyteである．

6.2 手 順

日本語の専門用語は一般用語で構成された複合名詞であることが多い．そこで，KWICレコードのキーを検索文字列にして一般用語辞書の検索を行い，文頭側の文脈文字列とキーまたは文末側の文脈文字列とキーにまたがる一般用語が存在する場合には，切断の誤りがあるものとして修正する．その結果，キー全体が一つの一般用語と一致するならば，その KWIC レコー

ドを削除する．一般用語辞書の検索は，文頭側からと文末側からの二方向について，最長一致優先の原則に従って行い，途中で行き詰まったときには一文字の読み飛ばしをする．なお，この処理に先立ちキーの末尾が'的'である KWIC レコードはすべて削除する．

切断誤りを回復するための基本的方法は，キーの文字列を含む最小の単語列を求めることである．この方法で，図2に示した例の(1)と(2)は文頭側から右向きに行う検索で，(3)は文末側から左向きに行う検索で誤りを回復することができる．しかし，(4)の場合は左向きの検索を行っても"はめ込み理論"しか抽出することができない．一般に，名詞列の先頭または末尾が仮名表記のために生じた切断の誤りは基本的方法で回復できるが，名詞列の途中が仮名表記のために生じた誤りは，完全に回復できないことがある．この場合，先頭の漢字列を含む文字列は用言とみなされてキーにならないこともあるが，本来名詞が存在しているのだから末尾の漢字列は必ず名詞と解析されるキーになる．そのため，左向きの検索の場合には，キーと文頭側文脈文字列にまたがる単語が存在するときは，続けて左向きに単語を検索するという方法をとっている(アルゴリズムの step. 6).

次に，このアルゴリズムを記述するために幾つかの記述上の約束と関数の定義を行う．

(i)　文字列 w の長さを $|w|$ で表す．

(ii)　二つの文字列 v と w の連結を $v\cdot w$ で表す．

(iii)　文字列 v が文字列 w の最左部分列 (prefix) であることを $v \fallingdotseq w$ で表す．

(iv)　正引き辞書の見出し文字列の集合を D_l で表し，逆引き辞書の見出し文字列の集合を D_r で表す．

(v)　文字列 s と文字列の集合 D に対して，$w \fallingdotseq s$, $w \in D$ を満たすすべての w における $|w|$ の最大値を $\max(s, D)$ で表す．

(vi)　文字列 s の末尾で終わる長さ i の部分列を $\mathrm{right}(s, i)$ で表し，先頭から始まる長さ i の部分列を $\mathrm{left}(s, i)$ で表す．

(vii)　文字列 s の並びを左右逆にしてできる文字列を $\mathrm{reverse}(s)$ で表す．

ここで，空文字列 $\varepsilon(|\varepsilon|=0)$ は任意の文字列 s と文字集合 D について，$\varepsilon \fallingdotseq s$, $\varepsilon \in D$ である．以上の記号と関数を用いて，一つの KWIC レコードに対して一般用語の除去と切断誤りの回復を行うアルゴリズムを次に示す．なお，記述中 KWIC レコードの文頭側の文脈文字列を C_l で，文末側の文脈文字列を C_r で，キーの文字列を key で表す．

〔一般用語の除去と切断誤り回復のアルゴリズム〕

Step. 1 (右向き走査の初期設定)
　$s \leftarrow \mathrm{key} \cdot C_r$.
　$p \leftarrow 0$.
$p \geqq |\mathrm{key}|$ になるまで Step. 2 を繰り返す．
Step. 2 (右向き走査)
　if $\max(s, D_l)=0$　then $p \leftarrow p+1$.
　　else $p \leftarrow p+\max(s, D_l)$.
　$s \leftarrow \mathrm{right}(s, |\mathrm{key} \cdot C_r| - p)$.
Step. 3 (切断誤りの回復)
　if $p > |\mathrm{key}|$ then　$\mathrm{key} \leftarrow \mathrm{left}(\mathrm{key} \cdot C_r, p)$,
　　$C_r \leftarrow s$.
Step. 4 (左向き走査の初期設定)
　$s \leftarrow \mathrm{reverse}(C_l \cdot \mathrm{key})$.
　$p \leftarrow 0$.
　count$\leftarrow$0.
$p \geqq |\mathrm{key}|$ になるまで Step. 5 を繰り返す．
Step. 5 (左向き走査1)
　if $\max(s, D_r)=0$　then $p \leftarrow p+1$.
　　else $p \leftarrow p+\max(s, D_r)$.
　$s \leftarrow \mathrm{right}(s, |C_l \cdot \mathrm{key}| - p)$.
　count$\leftarrow$count$+1$.
$p = |\mathrm{key}|$ ならば Step. 8 へ．
$p > |\mathrm{key}|$ ならば $\max(s, D_r) \neq 0$ のあいだ Step. 6 を繰り返す．
Step. 6 (左向き走査2)
　$p \leftarrow p+\max(s, D_r)$.
　$s \leftarrow \mathrm{right}(s, |C_l \cdot \mathrm{key}| - p)$.
　count$\leftarrow$count$+1$.
Step. 7 (切断誤りの回復)
　if $p > |\mathrm{key}|$ then　$\mathrm{key} \leftarrow \mathrm{right}(C_l \cdot \mathrm{key}, p)$,
　　$C_r \leftarrow \mathrm{reverse}(s)$.
Step. 8 (一般用語の除去)
count$=1$ ならば KWIC レコードを削除する．
〔アルゴリズム終〕

6.3 専門用語の抽出率

4章で述べた名詞の抽出実験と同一の入力データに対して行った，専門用語の抽出実験の結果を次に示す．

専門用語の総数 (a)	1403 個
抽出できなかった専門用語の個数 (b)	6 個
誤って抽出した文字列の個数 (c)	23 個

部分的に抽出した専門用語の個数（d）　9 個
余分な文字列を伴って抽出した
　文字列の個数（e）　9 個
ここで，
抽出率$=(a-b-d-e)/a$　　$\{(100-94.5)-(100-98.3)\}/(100-94.5)=69.1$ (%)
抽出の精度$=(a-b-d-e)/(a-b+c)$　　$\{(100-91.8)-(100-97.1)\}/(100-91.8)=64.6$ (%)
とすると，専門用語の抽出率は 98.3%，抽出の精度は 97.1% である．ここで接辞の除去を行わなかった場合，表 4 より $c=47$, $e=62$ となり，抽出率は 94.5%，抽出の精度は 91.8% となる．したがって，接辞の除去による抽出率および，抽出の精度の改善率は，

$$\{(100-94.5)-(100-98.3)\}/(100-94.5)=69.1 \quad (\%)$$

$$\{(100-91.8)-(100-97.1)\}/(100-91.8)=64.6 \quad (\%)$$

である．

専門用語抽出の精度が 100% でないため，実際の収集ではこの段階で人間による用語の選択操作が必要である．このとき，KWIC レコードのキーとして部分的にでも抽出されている専門用語は，人間が介入すること

81　　E82060018　　1

2つのはん関数の順序積の平均値。
Kuboの順序指数（J. Phys. Soc. Japan, 1962, 17）の自然な一般化である順序付けられたはん関数を導入。
交換する確率過程の2つのはん関数の積の平均に関する周知の定理（特にFurutsu-Novikovの定理）を非交換確率過程の場合に一般化。
閉じた量子系のじょう乱解析での応用を示す。
応答の高次平均が順序キュミュラントで表現され，物理系のGauss特性を最大限利用できる。

---------- FROM TITLE ----------

2つの　　はん関数
　　はんかんすう
2つのはん関数の　　順序積
　　じゅんじょせき

---------- FROM ABSTRACT ----------

する周知の定理（特に　　Furutsu-Novikov
　　Furutsu-Novikov
で表現され，物理系の　　Gauss特性
　　Gaussとくせい
Kuboの順序指数（　　J. Phys. Soc. Japan
　　J. Phys. Soc. Japan
　　Kubo
　　Kubo
閉じた量子系の　　じょう　乱解析
　　じょうらんかいせき
する確率過程の2つの　　はん関数
　　はんかんすう
である順序付けられた　　はん関数
　　はんかんすう
交換する　　確率過程
　　かくりつかてい
応答の　　高次平均
　　こうじへいきん
応答の高次平均が　　順序キュミュラント
　　じゅんじょきゅみゅらんと
Kuboの　　順序指数
　　じゅんじょしすう
程の2つのはん関数の　　積
　　せき
vikovの定理）を　　非交換確率過程
　　ひこうかんかくりつかてい
閉じた　　量子系
　　りょうしけい

の順序積の平均値。

の平均値。

の定理）を非交換確率過程の場合に一般化。

を最大限利用できる。

，1962，17）の自然な一般化である順

の順序指数（J. Phys. Soc. Jap

での応用を示す。

の積の平均に関する周知の定理（特にFur

を導入。

の2つのはん関数の積の平均に関する周知の

が順序キュミュラントで表現され，物理系の

で表現され，物理系のGauss特性を最大

（J. Phys. Soc. Japan, 19

の平均に関する周知の定理（特にFurut

の場合に一般化。

のじょう乱解析での応用を示す。

図 3　出　力　例
Fig. 3　Example of output.

とにより抽出できる．このような KWIC レコードも正しく抽出されたものとして扱うと，

抽出率$=(a-b)/a$

抽出の精度$=(a-b)/(a-b+c)$

となり，抽出率は 99.6%，抽出の精度は 98.4% となる．

まったく抽出できなかった 6 個の専門用語を次に示す．下線部の英字はその部分に対して与えられた解析単位を示している．

（1） 開放共振器近くの
KKKKKKGF

（2） 斜め入射の
KGKKF

（3） 飛しょう
K M G

（4） 部分的線形近似がえられる
KKKKKKKMG F

（5） 飽和蒸気圧にて
KKKKKGF

（6） 巨大電子なだれを
KKKKG J F

（5）を抽出できなかった原因は，付属語辞書に助詞“にて”が未登録であったことであり，（1）～（4），（6）は用言と解析されたために抽出できなかった．また，誤って解析した文字列は次のものである．

（1） 活用語尾をもたない上一段，下一段動詞の未然形と連用形．

（2） 用言の連用形またはそれに一般用語が連接したもの．

（3） 数式．

部分的にしか抽出できなかった専門用語は“磁気しゃへい”のように後半部分が仮名書きされ，なおかつその単語が一般用語辞書に未登録であったものである．余分な文字列を伴って抽出された専門用語は，“対応し波動場”のように専門用語の前に動詞の連用形が連接したものであった．

6.4 専門用語の読み振り

収集した専門用語を辞書的順序にソートするために，一般用語の除去と切断誤りの回復を行う過程で専門用語の候補に読みを振る．このとき，専門用語を構成している一般用語には一般用語辞書から得られる読みを振り，漢字（解析単位の種類が K になっているもの）には，漢字とその音読みを一対一に対応させた漢字の音読みテーブルを参照して読みを振っている．この方法で抽出された漢字表記の専門用語の 91.7% に正しい読みを振ることができた．なお，読み振りについては，より高い精度で読み振りを行う研究が報告されており[14)~16)]，今後改良の余地が残されている．

図 3 に一文献に対する出力結果を示す．キー中のスペースは，その位置で切断誤りの回復が行われたことを示している．

7. むすび

本論文で報告したシステムは，九州大学大型計算機センターの FACOM-M 382 上に PL/I を用いて実現した．この環境で，一文献の表題と抄録文を処理するのに要した時間は平均 326.3 msec であった．一文献の表題と抄録文の平均文字数は 174.4 文字である．参考に図 3 に示した結果を求めるのに要した時間は 391 msec である．このうち，右向きの走査に 101 msec 左向きの走査に 91 msec を要している．参考文献 9) の形態素解析をそのまま用いて専門用語を抽出する場合との精度の比較は行っていないが，処理時間は 4 分の 1 程度になっている．6 章のアルゴリズムを，単に，

$\max(\text{key}\cdot C_r, D_l)\geqq|\text{key}|$ ならば KWIC レコードを削除する．

と変更して，切断誤りの回復を行わない場合，処理時間は約 56% に短縮できるが，部分的に抽出する専門用語の個数は 43 個になり，抽出率は 95.9%，抽出の精度は 94.7% となる．したがって，切断誤りの回復処理による抽出率の改善率は，

$$\{(100-95.9)-(100-98.3)\}/(100-95.9)=58.5\ (\%)$$

である．

専門用語の抽出率，抽出の精度については，特殊自立語辞書，一般用語辞書を整備することにより，さらに高めることが可能である．形態素解析におけるコストの値についても今後，大量の実験により，最適値を決定しなければならない．また，機械翻訳で利用するための対訳の専門用語辞書を作成するためには，日本語と英語の表題，抄録文を同時に処理する方法[8)]を取り入れる必要がある．

謝辞 システムの作成に協力していただいた山下明男（現富士ゼロックス），古城慎太郎（現富士ゼロックス）の両氏，および有益なご助言をいただいた福岡大学首藤公昭教授に感謝の意を表します．なお，本研究の一部は，文部省特定研究「言語の標準化」による

ものである.

参考文献

1) 斉藤秀紀：電子計算機と漢テレによる用語総索引の作成，電子計算機による国語研究，国立国語研究所報告，No. 31, pp. 91-103 (1968).
2) 土屋信一：カナ入力による日本語文総索引の作成，電子計算機による国語研究Ⅳ，国立国語研究所報告，No. 46, pp. 35-43 (1972).
3) 靍岡昭夫：言語研究のための索引作成システム，電子計算機による国語研究Ⅷ，国立国語研究所報告，No. 59, pp. 1-17 (1976).
4) 中野　洋：索引作成のためのプログラムライブラリ，電子計算機による国語研究Ⅷ，国立国語研究所報告，No. 59, pp. 18-62 (1976).
5) 植村俊亮：電子計算機による自動索引の研究(上)，電子技術総合研究所報告，No. 743 (1974).
6) 植村俊亮：電子計算機による自動索引の研究(下)，電子技術総合研究所報告，No. 747 (1974).
7) 長尾　真，水谷幹男，池田浩之：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, pp. 110-117 (1976).
8) 荒木啓介，金子明夫，高野文雄，日夏健一：日本語文論文タイトルからのキーワード自動抽出システム (JAKAS)，情報処理学会自然言語処理研究会資料，26-3 (1981).
9) 吉村賢治，日高　達，吉田　将：未登録語を含む日本語文の形態素解析アルゴリズム，九州大学工学集報，Vol. 55, No. 6, pp. 635-640 (1982).
10) 吉村賢治，日高　達，吉田　将：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol. 24, No. 1, pp. 40-46 (1983).
11) 吉田　将，日高　達，稲永紘之，田中武美，吉村賢治：公用データベース日本語単語辞書の使用について，九州大学大型計算機センター広報，Vol. 16, No. 4, pp. 335-361 (1983).
12) 首藤公昭：文節構造モデルによる日本語の機械処理に関する研究，福岡大学研究所報，第 45 号 (1980).
13) 日高　達，吉田　将，稲永紘之：拡張 B-tree と日本語単語辞書への応用，電子通信学会論文誌，Vol. J 67 D, No. 4, pp. 399-404 (1984).
14) Knuth, D. E.: *The Art of Computer Programming*, Vol. 3, Addison-Wesley Publishing Company, Reading, Massachusetts (1973).
15) 荒木啓介，板山和彦：JICST の実用的全自動漢字—カナ変換システム，K-KACS について，情報処理，Vol. 20, No. 10, pp. 917-923 (1979).
16) 藤崎哲之助：動的計画法による漢字仮名混り文の単位切りと仮名ふり，情報処理学会自然言語処理研究会資料，28-5 (1981).
17) 宮崎正弘，白井　論，大山芳史，後藤滋樹，池原　悟：日本文音声出力のための言語処理，情報処理学会，自然言語処理技術シンポジウム論文集，pp. 5-16 (1983).
18) 田中康仁：専門用語の自動抽出，第 17 回情報科学技術研究集会論文集，pp. 119-125 (1980).

(昭和 60 年 3 月 4 日受付)
(昭和 60 年 7 月 18 日採録)