自然言語処理　42－1
（１９８４．３．１６）

専門用語の自動収集システムについて

吉村賢治・山下明男・日高　達・吉田　将
（九州大学　工学部）

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，専門用語辞書の開発が望まれている．しかし，日本語における専門用語の大部分は複合語と外来語で，その数が膨大であること，新しい用語の出現数が多いことなどから，専門用語を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．筆者らは，現在、科学技術文献の表題と抄録文から専門用語を収集するシステムを開発している。この専門用語の自動収集システムは、次のようなサブ・システムで構成されるが、本稿では実験システムが完成している専門用語の自動抽出について報告する。

(1) 専門用語の自動抽出
(2) 専門用語辞書の作成・管理
(3) シソーラスの作成・管理

専門用語を表題，抄録文等から自動収集するための手法としては，不要語(stop word) を除去する方法が一般的であり，実用性も高いと考えられる．不要語除去の具体的方法として次の二つの方法がある。

(1) 不要語テーブルを用いて文字列のパターン・マッチングを行なう方法[1]。
(2) 単語辞書と文法規則のテーブルを用いて文法解析を行なう方法。

方法(1) は文法解析を行なわないため処理は簡単であるが、不要語テーブルを作成するために大量のデータを調査しなければならない。また、不要語テーブルとのマッチングだけで不要語を除去した場合、平仮名一文字で表わされる助詞などが原因で誤った切断が生じる。これを避けて専門用語抽出の精度を上げるためには、不要語テーブルとのマッチングで生じる切断の誤りを抑制するために切断禁止ルールのテーブル等が必要となり、テーブル間の干渉の管理問題などが発生する。

一方、方法(2) には上記のような問題はないが、文法解析を行なうために処理が複雑になり、大規模な自立語辞書を使用した場合、処理時間も長くなる。また、一般に大部分の専門用語はシステムの単語辞書に登録されていないため、未登録語を含んだ日本語文を処理できる文法解析システムが必要である。

本稿で報告する専門用語の自動収集システムは日本語文の形態素解析システム[2]を利用している。この形態素解析システムは未登録語を含む日本語文の解析が可能であり、この特徴を利用して上に述べた大規模な自立語辞書の検索に要する時間の削減を図っている。

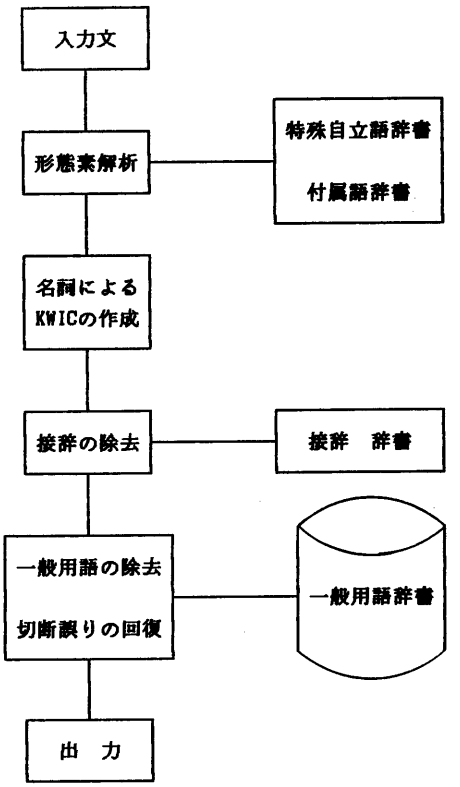
専門用語の自動抽出システムは図－１に示すモジュールから構成されている。以下、図－１の各モジュールについて説明し、日本科学技術情報センターの科学技術文献速報・電気工学編VOL.25,NO.6 から抽出した表題及び抄録文に対して行なった実験とその結果について述べる。なお、ここで抽出の対象としている専門用語の品詞は名詞だけである。

２．形態素解析

２．１　形態素解析の手順

このシステムでは、参考文献(2) で報告した形態素解析アルゴリズムを用いている。本章ではアルゴリズムの概略、及び一般の形態素解析システムとの相違点について述べる。

この形態素解析アルゴリズムでは、まず入力文を文頭から文末に走査し、入力文を構成しているすべての解析単位（入力文が分解される解析の単位。一般には単語であるが、以下では解析単位と呼ぶ。）について、　５項



図－１　システムの構成

組（i,j,α,k,a）を作成する。この５項組をアイテムと呼び、一つの入力文に対して作成されるアイテムの集合をパーズリストと呼ぶ。ここで i、j は入力文における解析単位の始点と終点、αは解析単位の文法情報である。専門用語の自動収集システムにおける解析単位はその素性によって表－１に示す１０種類に分類される。a はその種類を示し、表－１の右の記号を値として持つ。また、解析単位にはその種類に応じてコストが設定されており、k は文頭からその解析単位までのコストの総和である。このコストの総和は連接する解析単位が文法的な連接規則をみたしている列についてのみ求める。

表－１において、（vi）～（viii）が未登録語に対処するための解析単位であり、漢字仮名混りの未登録語は（vi）と（viii）で、外来語の未登録語は（vii）で処理される。また、連結記号とは文字列を連結して専門用語を構成するハイフン等の記号である。

ここで、解析単位の種類を表わす記号を添字にして各解析単位のコストを C_J、C_N、C_F、C_G、C_H、C_K、C_R、C_M、C_C、C_S で表わす。筆者らは参考文献(3) で もっともらしい解析結果から抽出するためのヒューリスティックスとして文節数最小法を提案した。解析結果のコストの総和が小さいものから出力するとき、このヒューリスティックスに従うために

C_J，C_N，$C_H > C_F$，C_G

とする。また、ほとんどの表記は正書法に従っており、本来の未登録語は稀にしか出現しないという仮定から、

$C_H > C_J$，C_N

C_K，C_R，$C_M > C_J$，C_N

とする。以上のことと字種の違いを考慮して、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。

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

なお、本稿で報告する実験は、

C_C，$C_S = 0$

$C_F = 1$

$C_G = 2$

	解析単位	記号
(i)	正書表記の自立語	J
(ii)	数詞	N
(iii)	付属語	F
(iv)	活用語尾	G
(v)	非正書表記の自立語	H
(vi)	漢字	K
(vii)	アルファベット・片仮名列	R
(viii)	平仮名	M
(ix)	連結記号	C
(x)	記号	S

表－１　解析単位の種類

C_J，$C_N = 4$

C_H，C_K，$C_R = 6$

$C_M = 14$

で行なったものである。

形態素解析システムでは、作成されたパーズリストを文末側から文頭側に走査して、入力文の解析結果を抽出するが、専門用語の自動収集システムでは３章で述べる手順に従ってパーズリストから名詞（列）をキーとした KWIC レコードを作成する。

２．２　特殊自立語辞書

参考文献(2) の形態素解析システムでは単語数約83,000、見出し語数約170,00の自立語辞書[4]を使用しているが、１章で述べたように文法解析を行なう専門用語の自動収集システムでは二次記憶上に置かれた自立語辞書の検索時間が問題になる。そこで、この専門用語の自動収集システムにおける形態素解析では、上記の辞書の代わりに主記憶上に常駐することができる小規模の自立語辞書（これを特殊自立語辞書と呼ぶ）を使用している。表－２に示すように、特殊自立語辞書には、連体詞、接続詞、副詞、形容詞など専門用語やその構成要素でない814単語が登録されている。なお、特殊自立語辞書の作成にあたっては首藤のデータ[5]を参考にしており、多くの慣用句的な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して出力される。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、形態素解析では漢字（K）と同様に扱う。

３．名詞によるKWICの作成

パーズリストからはコストの総和が最小となる解析結果のみを抽出する。この解析結果抽出の過程で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。いま、解析単位の種類を示す記号でそれぞれの解析単位を表わすと、KWICレコードのキーになる解析単位の列Xは次のように記述される。

品　詞	単語数
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

表－２　特殊自立語辞書の内容

X ::= {K, M, R}
X ::= X・X
X ::= X・C・X
X ::= N・C・X
X ::= N・X
X ::= X・G*

ここで、括弧"{"、"}"は、その中の一つを選択することを表わし、中点"・"は連結を表わす。また、G*は5段動詞の連用形の活用語尾である。

KWICレコードの作成は文献単位に行なう。一つの文献のKWICレコードは、表題から作成されたものと抄録文から作成されたものに分け、4章で述べる接辞の除去、5章で述べる一般用語の除去と切断誤りの回復を行なった後、それぞれをキーの部分でソートして表示する。KWICレコードを作成した段階の出力例を図－2に示す。

この段階で抽出され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（列）である。ここでは、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文献に対して行なった抽出実験の結果を示す。

対象：日本科学技術情報センター・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文
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

いま、抽出されるべきものの個数をa、抽出されるべきものの中で抽出されなかったものの個数をb、抽出されるべきもの以外で抽出されたものの個数をcとして、

抽出率＝（a－b）／a

抽出の精度＝（a－b）／（a－b＋c）

と定義すると、KWICレコードの作成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殆どのものが、図－3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。また、誤って抽出されたcに相当するものは、殆どが形容動詞の語幹であった。

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
2つのはん関数の	順序積
2つのはん関数の順序積の	平均値

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

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

の順序積の平均値。
の平均値。
。

の順序指数（J. Phys. Soc. Jap
（J. Phys. Soc. Japan, 19
, 1962, 17）の自然な一般化である順
である順序付けられたはん関数を導入。
を導入。
。
の2つのはん関数の積の平均に関する周知の
の積の平均に関する周知の定理（特にFur
の平均に関する周知の定理（特にFurut
に関する周知の定理（特にFurutsu-
の定理（特にFurutsu-Noviko
（特にFurutsu-Novikovの定
の定理）を非交換確率過程の場合に一般化。
）を非交換確率過程の場合に一般化。
の場合に一般化。
。
のじょう乱解析での応用を示す。
での応用を示す。
を示す。
の高次平均が順序キュミュラントで表現され
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
のGauss特性を最大限利用できる。
を最大限利用できる。
できる。

図－2 名詞によるKWICの作成結果

4．接辞の除去

2章で述べた形態素解析では接辞処理を行なっていないため、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。そこで、表－3に示す接頭語辞書、接尾語辞書を用いて、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各々移動する。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、この誤りの殆どは次のステップで回復される。図－2に示したKWICレコードのキーから接辞を除去した後の出力例を図－4に示す。

5．一般用語の除去と切断誤りの回復

このステップまでに作成されたKWICレコードの多くは一般用語がキーになっている。このステップ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なう。また、形態素解析の段階で生じた専門用語の切断誤りの回復も同時に行なう。

現在、一般用語辞書としては、参考文献(2)の形態素解析システムで使用している自立語辞書から正書表記の名詞だけを抽出して構成したものを使用している。この一般用語辞書の見出し語数は79,102個で、処理の都合により、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を用意している。辞書のデ

磁気しゃ▲へ▲い
二乗平均平方根的広▲が▲り
じょう▲乱解析
不変はめ▲込み理論

図－3　抽出を誤った名詞の例

	接頭語	接尾語
1	各	やすい
2	諸	下
3	当該	以外
4	本	間
5	両	時
6		上
7		側
8		全体
9		中
10		等
11		内
12		後半
13		分
14		用

表－3　接辞辞書

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

2つの	はん関数
2つのはん関数の	順序積
2つのはん関数の順序積の	平均値

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

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

の順序積の平均値。
の平均値。
。

の順序指数（Ｊ．Ｐｈｙｓ．Ｓｏｃ．Ｊａｐ
（Ｊ．Ｐｈｙｓ．Ｓｏｃ．Ｊａｐａｎ，１９
，１９６２，１７）の自然な一般化である順
である順序付けられたはん関数を導入。
を導入。
。
の２つのはん関数の積の平均に関する周知の
の積の平均に関する周知の定理（特にＦｕｒ
の平均に関する周知の定理（特にＦｕｒｕｔ
に関する周知の定理（特にＦｕｒｕｔｓｕ－
の定理（特にＦｕｒｕｔｓｕ－Ｎｏｖｉｋｏ
（特にＦｕｒｕｔｓｕ－Ｎｏｖｉｋｏｖの定
の定理）を非交換確率過程の場合に一般化。
）を非交換確率過程の場合に一般化。
の場合に一般化。
。
のじょう乱解析での応用を示す。
での応用を示す。
用を示す。
の高次平均が順序キュミュラントで表現され
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
のＧａｕｓｓ特性を最大限利用できる。
を最大限利用できる。
用できる。

図－4　接辞除去の結果

ータ構造としては拡張B-treeとTRIEを併用しており[6]、容量は正引き辞書が1,121Kbyte、逆引き辞書が1,178Kbyteである。

5．1　アルゴリズム

一般用語の除去と切断誤り回復のアルゴリズムを簡潔に記述するために、ここで幾つか記述上の約束と関数の定義を行なう。

（ｉ）文字列ｗの長さ（文字数）を｜ｗ｜で表わす。
（ii）二つの文字列ｖ，ｗの連結をｖ・ｗで表わす。
（iii）文字列ｖが文字列ｗの最左部分列(prefix)であることをｖ≒ｗで表わす。
（iv）正引き一般用語辞書の見出し文字列を要素とする集合をD_Lで表わし、逆引き一般用語辞書の見出し文字列を要素とする集合をD_Rで表わす。
（ｖ）文字列ｓと文字列の集合Ｄに対して、ｗ≒ｓとｗ∈Ｄを満たす総ての｜ｗ｜における最大値をＭａｘ（ｓ，Ｄ）で表わす。
（vi）文字列ｓにおける右側から長さｉの部分文字列をＲｉｇｈｔ（ｓ，ｉ）で表わし、左側から長さｉの部分文字列をＬｅｆｔ（ｓ，ｉ）で表わす。
（vii）文字列ｓの並びを左右逆にしてできる文字列をＲｅｖｅｒｓｅ（ｓ）で表わす。

ここで、空文字列ε（｜ε｜＝０）は任意の文字列ｓと文字列の集合Ｄについて、ε≒ｓ、ε∈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なうアルゴリズムを右に示す。

右に示すアルゴリズムを総てのKWICレコードに対し適用して、一般用語の除去と切断誤りの回復を行なう。なお、この処理に先立ち、キーの末尾が”的”であるKWICレコードは総て除去しておく。また、アルゴリズムの記述中、KWICレコードにおける前の文脈文字列をC_Lで、後の文脈文字列をC_Rで、キーの文字列をＫｅｙで表わす。

図－４に示した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ない、キーで昇順にソートした結果を図－５に示す。図－５において”【”と”】”で囲まれたキーは、一般用語として除去されたキーであり、キーの文字列中にスペースを含むキーは、その位置で切断誤りの回復が行なわれたキーである。

一般用語の除去と切断誤りの回復を行なった時点で、専門用語の抽出を完了する。３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なった、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専門用語の個数	e ・・・	9個

[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←Ｋｅｙ・C_R．
　ｐ←０．

ｐ≧｜Ｋｅｙ｜になるまでＳｔｅｐ．２を繰り返す。

Ｓｔｅｐ．２　（右向き走査）
　Ｍａｘ（ｓ，D_L）＝０　ならば　ｐ←ｐ＋１．
　Ｍａｘ（ｓ，D_L）≒０　ならば
　　ｐ←ｐ＋Ｍａｘ（ｓ，D_L）．
　ｓ←Ｒｉｇｈｔ（ｓ，｜Ｋｅｙ・C_R｜－ｐ）．

Ｓｔｅｐ．３　（切断誤りの回復）
　ｐ＞｜Ｋｅｙ｜　ならば
　　Ｋｅｙ←Ｌｅｆｔ（Ｋｅｙ・C_R，ｐ）
　　C_R←ｓ．

Ｓｔｅｐ．４　（左向き走査の初期設定）
　ｓ←Ｒｅｖｅｒｓｅ（C_L・Ｋｅｙ）．
　ｐ←０．
　Ｃｏｕｎｔ←０．

ｐ≧｜Ｋｅｙ｜になるまでＳｔｅｐ．５を繰り返す。

Ｓｔｅｐ．５　（左向き走査１）
　Ｍａｘ（ｓ，D_R）＝０　ならば　ｐ←ｐ＋１．
　Ｍａｘ（ｓ，D_R）≒０　ならば
　　ｐ←ｐ＋Ｍａｘ（ｓ，D_R）．
　ｓ←Ｒｉｇｈｔ（ｓ，｜C_L・Ｋｅｙ｜－ｐ）．
　Ｃｏｕｎｔ←Ｃｏｕｎｔ＋１．

Ｍａｘ（ｓ，D_R）≒０のあいだＳｔｅｐ．６を繰り返す。

Ｓｔｅｐ．６　（左向き走査２）
　ｐ←ｐ＋Ｍａｘ（ｓ，D_R）．
　ｓ←Ｒｉｇｈｔ（ｓ，｜C_L・Ｋｅｙ｜－ｐ）．
　Ｃｏｕｎｔ←Ｃｏｕｎｔ＋１．

Ｓｔｅｐ．７　（切断誤りの回復）
　ｐ＞｜Ｋｅｙ｜　ならば
　　Ｋｅｙ←Ｒｉｇｈｔ（C_L・Ｋｅｙ，ｐ）
　　C_R←Ｒｅｖｅｒｓｅ（ｓ）．

Ｓｔｅｐ．８　（一般用語の除去）
　Ｃｏｕｎｔ＝１ならば、KWICレコードを除去する。

[アルゴリズム終わり]

ここで、

抽出率＝（ａ－ｂ－ｄ－ｅ）／ａ

抽出の精度＝（ａ－ｂ－ｄ－ｅ）／（ａ－ｂ＋ｃ）

とすると、専門用語の抽出率は 98.3%、専門用語抽出の精度は 97.1%である。専門用語抽出の精度が100%でないため、この段階で人間による用語の選択作業が必要になる。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により正しく抽出することができる。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%になる。

全く抽出できなかった専門用語は次のものである。

(1) 開放共振器近くの
(2) 斜め入射の
(3) 飛しょう
(4) 部分的線形近似がえられるが
(5) 超和蒸気圧にて
(6) 巨大電子なだれを

これらのうち、(5) を抽出できなかった原因は付属語辞書の不備である。

本稿で報告した実験システムは、九州大学大型計算機センターの FACOM M382 上にPL/1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する時間の平均は1439.3msec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－５に示した結果を求めるのに要した時間は1585msecである。この内、右向きの走査に要した時間は 696msecで、左向きの走査に要した時間は 779msecであった。

６．あとがき

本稿では形態素解析システムを利用した専門用語の自動抽出システムについて報告した。専門用語抽出の精度については、特殊自立語辞書、一般用語辞書のデータを整備することにより、さらに高めることが可能である。

大規模な自立語辞書を用いた形態素解析を利用して専門用語を抽出する場合に対する、抽出の精度及び処理時間の比較はまだ行なっていないが、処理時間は殆ど改善されていないようである。この原因は切断誤りの回復処理において、二重の辞書引きを行なうことにある。５章のアルゴリズムを、

$s \leftarrow Key \cdot C_R$,

$Max(s, D_L) \geq |Key|$ ならば

KWICレコードを除去する。

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

2つの	はん関数
2つのはん関数の	順序積
2つのはん関数の順序積の	【平均値】

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

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

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。
。
である順序付けられたはん関数を導入。
の高次平均が順序キュミュラントで表現され
を示す。
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
の定理（特にＦｕｒｕｔｓｕ－Ｎｏｖｉｋｏ
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
（特にＦｕｒｕｔｓｕ－Ｎｏｖｉｋｏｖの定
）を非交換確率過程の場合に一般化。
。
の場合に一般化。
のＧａｕｓｓ特性を最大限利用できる。
に関する周知の定理（特にＦｕｒｕｔｓｕ－
できる。
のじょう乱解析での応用を示す。

図－５　一般用語の除去と切断誤り回復の結果

と変更して、切断誤りの回復処理を行なわない場合、処理時間は 100文献で平均して一文献あたり 601.9msecになるが、部分的に抽出した専門用語の個数は４３個になり、

抽出率 (a-b-d-e)/a = 95.9%

抽出の精度 (a-b-d-e)/(a-b+c) = 94.7%

となる。従って、切断誤りの回復処理による抽出率の改善率は (4.1-1.7)/4.1 より 58.5% である。

本研究は昭和５８年度文部省科学研究費特定研究「言語の標準化」によった。

７．参考文献

(1) 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。

(2) 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、第55巻、第 6号、1982。

(3) 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、第24巻、第 1号、1983。

(4) 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、第16巻、第 4号、1983。

(5) 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。

(6) 日高達・吉田将・稲永紘之：”拡張B-treeによる日本語単語辞書の作成”、情報処理学会自然言語処理研究会資料33-8、1982。