漢字の出現頻度特性を用いた 日本語文献の機械処理

細野公男^{*1}, 原田隆史^{*1}, 梅田茂樹^{*2}
諸橋正幸^{*2}, 後藤智範^{*3}, 守屋 智^{*4}

〔抄 録〕 文献情報を有効に活用するには、個々の文献に対してキーワード、分類コード等の情報を的確かつ迅速に付与する必要がある。この要件を達成する一つ的手段として、キーワードの自動抽出、文献の自動分類が考えられる。これらに関する研究の多くは、文献中出现する単語を手がかりにするものである。しかし、筆者らは、日本語文献の場合そのキーワードの多くが漢字もしくはカタカナ表現されることに注目し、単語ではなく文献中出现する漢字を手がかりに、これらの情報を自動付与するいくつかの方法を考案し、実験を行ってきた。本稿はこれらの方法の特徴、実験成果を紹介し、あわせてその残された問題点を明らかにする。

1. は じ め に

新たな研究・開発を行うにあたっての情報源として、学術論文、技術レポート、特許などの記録資料が極めて重要であることは論を待たない。そのため、必要な情報を迅速に入手するための検索技術の開発や情報システムの構築が、いろいろ試みられている。

適切なキーワードの選択は、文献の分類および文献へのアクセス・ポイントの付与に当たって不可欠であるが、その作業には知的な処理が伴うので、多量の文献を対象とする場合には次の問題が生じる。

① 適切なキーワードの選択にはかなりの時間を

要するので、分類および索引作業が遅れがちになる。

② 高度の処理能力を持つ索引者を多数そろえることが困難であるため、選択されたキーワードの品質にかなりのバラツキが生じる恐れがある。

その結果、情報検索システムに蓄積される情報の鮮度が落ちるだけでなく、検索性能も低くなることが予想される。この種の問題を解決する一つ的手段として、コンピュータ技術を用いた自然語処理の導入がある。例えば、文献の標題あるいは抄録の中から、文献の主題を的確に表現すると期待される語をコンピュータで抽出することができれば、索引および分類作業を機械的に行うことが

* 1 HOSONO Kimio/HARADA Takashi 慶應義塾大学文学部 (〒108 港区三田 2-15-45)

* 2 UMEDA Shigeki/MOROHASHI Masayuki 日本アイ・ビー・エム(株) サイエンス・インスティテュート (〒102 千代田区三番町 5-19 第36興和ビル)

* 3 GOTO Tomonori 愛知淑徳大学 (〒480-11 愛知県愛知郡長久手町)

* 4 MORIYA Satoru 野村コンピュータ・システム(株) (〒103 中央区日本橋小網町 6-7 第2山万ビル)

可能となる。自動分類や自動索引はその顕著な例である。

コンピュータを使用する自然語処理の研究は、H. P. Luhn の KWIC 索引を嚆矢とすることができよう。その後 Luhn は、文献中の個々の語が文献中に出現する頻度に基づいてその語の重要度を考え、特に出現頻度が中程度の語が文献の主題を最もよく表現するとした¹⁾。

彼のアプローチは、文献中に出現する語の統計的な傾向に基づいて文献の主題を抽出しようとするものである。これと同様のアプローチをとるものとして、自動分類に Bayesian モデルを適用した Maron²⁾、条件付き確率の手法を導入した Hamill と Zamora³⁾、因子分析法を使用した Borko と Bernick⁴⁾、我が国では西村、岩坪^{5),6)}などの研究がある。

一方、自然語処理に統計的手法ではなく、構文上各語が持つ役割、語と語の間の関連についての情報を用いるアプローチもある。その 1 例がアメリカ DDC の機械補助索引である^{7),8)}。

単語の単位が明確である英語文献では、単語を処理単位として種々の機械処理が試みられている。一方、日本語文献では単語の単位が不明確であるので、単語を機械的に同定することが困難である。したがって、日本語文献の機械処理を行う場合には、前もって単語を切り出すなど手間のかかる前処理を必要とすることが多い。

こうした処理を回避するための 1 手段として、文字単位でなんらかの意味を表現することができる漢字を処理単位とみなして機械処理を行うことが考えられる。この場合、漢字 1 字が意味のレベルで英語の単語 1 語に相当するとみなすことができれば、英語等で行われてきた研究と同様のアプローチが簡単に適用できるようになる。

筆者らは、これまで JICST 科学技術データベース電気工学編磁気テープを使用して、上述の観点から重要漢字の決定、重要漢字に基づく文献の自動分類、キーワードの自動抽出に関して種々の実験を行ってきた。本稿はその成果を総括して、このようなアプローチの特色を紹介し、あわせて

その妥当性を明らかにするものである。実験の対象としたのは、JICST 1983 年度版磁気テープ 8 巻分である。

2. 漢字と主題分野との関連

一般に、文献中にある語または句を重要語として選定するための基準には、次の二つが考えられる。

- ① 主題表現力：その語句が文献の主題内容を構成している概念を表していること
- ② 主題識別力：その語句が文献の属する主題分野を明確に識別できる概念を表していること

この考え方を、そのまま漢字 1 字に適用すると、重要漢字とは主題表現力がある漢字、もしくは主題識別力のある漢字を意味することになる。筆者らは重要語の考え方を適用した重要漢字、また一般語の考え方を適用した一般漢字の抽出法を漢字 1 字単位で検討し、それに基づく実験を行った^{9),10)}。

実験に使用した「電気工学編」データは、JICST 統一分類コードに従って、表 1 に示す 12 の分野に分類されている。これにより、データベース中に出現するすべての漢字について各分野ごとに出現

表 1 JICST 統一分類コードに基づく
電気工学主題分野

No.	分類項目タイトル(小分野)	分類コード
1	計測学 計測機器	AD
2	電磁気学 光学	BD
3	電子物性 磁性 光物性	BM
4	生体工学	EL
5	システム・制御工学一般	IA
6	制御工学	IC
7	計算機方式 ハードウェア	JC
8	計算機利用技術	JE
9	電気工学一般	NA
10	電力工学	NB
11	電子工学	NC
12	通信工学	ND

	C ₁	C ₂		C _j		Total
K ₁	f ₁₁	f ₁₂		f _{1j}		Σf _{1j}
K ₂	f ₂₁	f ₂₂				Σf _{2j}
.....						
K _i	f _{i1}			f _{ij}		Σf _{ij}
.....						
K _m						
Total	Σf ₁₁	Σf ₁₂		Σf _{1j}		ΣΣf _{ij}

f_{ij} : 分野 C_j における漢字 K_i の出現頻度

図1 漢字の主題分野別出現頻度表

頻度をカウントし、図1に示す形式の主題分野別出現頻度表を作成し、実験の基礎データとした。

2.1 各主題分野ごとの重要漢字

ある漢字が特定の分野を代表する漢字であれば、特定の主題分野での出現率が、その分野を含む全体の平均出現率よりも大きくなるはずである。すなわち、図2で示すように平均出現率と特定の分野での出現率とを軸として、全漢字をプロットした図を考えたとき、図の斜線部にプロットされた漢字が、その特定の分野での重要漢字であるとする考え方である。

この考え方に従えば、各漢字の重要度は次のようなプロセスで求められる。

- i) 漢字 K_i の分野 C_j における出現率を、
 分野出現率 = $f_{ij} / \sum_i f_{ij} \dots (1)$

f_{ij} : 分野 C_j における漢字 K_i の出現頻度
 (マトリックスの各要素)

$\sum_i f_{ij}$: 分野 C_j における全漢字の出現頻度の和
 (各要素を縦に加えたもの)

また、漢字 K_i の全体での平均出現率を、
 分野平均出現率 = $\sum_j f_{ij} / \sum \sum f_{ij} \dots (2)$

$\sum_j f_{ij}$: 漢字 K_i の全分野での出現頻度の和
 (各要素を横に加えたもの)

$\sum \sum f_{ij}$: 全分野での全漢字の総出現頻度
 (全要素を加えたもの)

で表す。

- ii) 漢字 K_i の重要度を次式で求める。

$$\text{重要度} = \text{分野出現率} - \text{分野平均出現率} \dots (3)$$

- iii) そして重要度の値が十分大きい漢字 K_i を、分野 C_j での重要漢字の候補と考える。

このように決めると、全体での出現頻度は小さいが特定の分野にだけ集中する漢字は、重要漢字と見なされることになる。そこで、

- iv) 全体的に出現頻度の少ない漢字に対しては、分野出現率と分野平均出現率との差を、分野平均出現率で除した値(重要率)を重要漢字を決めるための指標とする。つまり、

$$\text{重要率} = \frac{\text{分野出現率} - \text{分野平均出現率}}{\text{分野平均出現率}} \dots (4)$$

(4)式のように重要漢字を定義すると、全体的な出現頻度は小さいが、特定の分野 C_j にだけ集中して出現する漢字の重要率は大きいものとなる。

筆者らは重要度および重要率を標題、抄録の両方を対象にして求めた。このうち、生体工学分野と電子工学分野において標題を対象に抽出した重要漢字を、重要率の降順に示したものが、表2、表3である。例えば表2における漢字‘血’の場合、分野出現率および分野平均出現率がそれぞれ9.76543, 0.4212であるので、(3)式と(4)式から重要度は9.34423, 重要率は22.18478となる。なお、表2および表3の重要度、重要率の値は、識別し

図2 漢字の分野出現率と
 分野平均出現率による分類

表 2 生体工学分野における重要度と重要率(標題)

	分野出現率	分野平均出現率	重 要 度	重 要 率
血	9.76543	0.42120	9.34423	22.18478
骨	5.55289	0.25828	5.29461	20.49949
臓	5.93585	0.27815	5.65770	20.34046
脈	4.11680	0.21854	3.89826	17.83774
弁	3.73384	0.21854	3.51530	16.08538
心	13.69076	1.16426	12.52650	10.75919
人	9.47821	0.90598	8.57223	9.46183
覚	4.69123	0.46491	4.22632	9.09061
臨	1.72331	0.19867	1.52464	7.67423
織	2.01053	0.25431	1.75622	6.90582
助	2.10627	0.27417	1.83210	6.68235
室	2.87218	0.41722	2.45496	5.88408
節	2.96792	0.46888	2.49904	5.32980
図	3.35088	0.62782	2.72306	4.33732
食	1.14887	0.21854	0.93033	4.25702
細	2.10627	0.42914	1.67713	3.90811
吸	2.20201	0.47285	1.72916	3.65688
診	1.91479	0.42120	1.49359	3.54603
両	0.95739	0.22252	0.73487	3.30248
工	9.09526	2.20534	6.88992	3.12419

表 4 分野出現偏差度 1 (標題)

	分野平均出現率	分野出現偏差度 1
御	5.47164	474.72995
制	6.26636	425.58094
血	0.42120	211.71238
測	6.40544	209.49173
識	1.14439	167.15651
光	9.37372	154.32549
電	27.97811	150.92364
心	1.16426	139.39392
認	0.90200	133.97842
像	2.25700	130.73752
臓	0.27815	118.14563
声	1.64904	116.68007
骨	0.25828	110.84091
画	2.78151	97.62393
信	7.11671	89.80859
定	10.30751	89.73587
人	0.90598	85.37648
温	1.93514	84.41867
肺	0.17881	80.12223
筋	0.17483	78.02106

表 3 電子工学分野における重要度と重要率(標題)

	分野出現率	分野平均出現率	重 要 度	重 要 率
池	6.28051	1.99077	4.28974	2.15481
倍	0.60224	0.19470	0.40754	2.09316
増	4.68887	1.57354	3.11533	1.97982
晶	4.68887	1.60533	3.08354	1.92081
幅	5.05452	1.79606	3.25846	1.81422
陥	0.70978	0.26225	0.44753	1.70650
欠	0.75280	0.28212	0.47068	1.66836
厚	1.13995	0.44106	0.69889	1.58456
太	6.45258	2.51528	3.93730	1.56535
陽	6.83973	2.67423	4.16550	1.55764
膜	5.87185	2.34839	3.52346	1.50037
陰	0.55922	0.22649	0.33273	1.46907
薄	3.20478	1.33910	1.86568	1.39323
純	0.81732	0.34172	0.47560	1.39178
振	5.78581	2.46363	3.32218	1.34848
半	5.03301	2.15369	2.87932	1.33692
演	0.92487	0.39735	0.52752	1.32759
積	4.58133	2.08216	2.49917	1.20027
路	12.43197	5.69019	6.74178	1.18480
酸	2.88215	1.31923	1.56292	1.18472

表 5 分野出現偏差度 2 (標題)

	分野平均出現率	分野出現偏差度 2
電	27.97811	3445.82100
光	9.37372	3363.81700
御	5.47164	2947.52900
測	6.40544	2944.12900
信	7.11671	2811.41000
制	6.26636	2740.22300
識	1.14439	2464.08600
血	0.42120	2259.64300
認	0.90200	1970.38700
像	2.25700	1945.59300
通	3.63981	1874.35600
発	9.17901	1743.87400
声	1.64904	1741.58900
心	1.16426	1512.68200
画	2.78151	1448.56500
力	8.16574	1432.24100
臓	0.27815	1269.31300
波	9.05185	1266.75900
衛	1.14439	1217.41600
定	10.30751	1209.31500

やすいように本来の値を 10^3 倍したものである。

電子工学分野においては重要度と重要率の差は顕著ではないが、生体工学分野においては、重要度では抽出されなかった重要漢字が、重要率では明確に抽出されている。これは、電子工学分野の全文献数が、生体工学分野の約 5 倍であることによる。

2.2 主題分野全体での重要漢字

前節では各主題分野ごとの重要漢字について議論したが、ここでは電気工学という学問分野全体にわたる重要漢字について考える。

一般的な漢字は、どの主題分野でも平均的に出現すると考えると、逆に重要漢字は、各主題分野間での出現のばらつきが大きくなるだろうと予想される。そこで、出現のばらつきを測る尺度(分野出現偏差度)を次の(5-1)、(5-2)で定義する。

$$\text{分野出現偏差度 1} = \frac{\sum_i (\text{分野出現率} - \text{分野平均出現率})^2}{\text{分野平均出現率}} \quad \dots(5-1)$$

$$\text{分野出現偏差度 2} = \frac{\sum_i (\text{分野出現頻度} - \text{分野平均出現頻度})^2}{\text{分野平均出現頻度}} \quad \dots(5-2)$$

このように決めると、これらの値が 0 に近い漢字は平均的な頻度傾向を示す。また、これらの値が大きい漢字は、分野出現傾向が平均的な頻度傾向と異なることになる。このとき、頻度そのものを用いるか、あるいは出現率を用いるかの違いで 2 通りの指標ができる。出現率を用いると各分野内の漢字の出現頻度を総出現頻度で基準化することになる。

一方、出現頻度を用いた時、分野出現偏差度 2 は統計学でいう χ^2 値と等価になる。出現率を用いたとき、分野出現偏差度 1 は検定には使えないが、値の大小を比較したとき、分野間のばらつきをある程度指標化することができる。標題および抄録を対象にしてすべての漢字について算出したもののうち、分野出現偏差度 1 および 2 の大きいものを表 4、5 にあげた。

頻度が高く分野出現偏差度 2 の小さい漢字は、

どの分野にも出現する一般的な漢字であるといえる。この例を表 6 にあげる。

3. 漢字の出現頻度を利用した自動分類

既述したように、漢字 1 文字と主題分野との間には、何らかの主題的な関連性があると考えられる。そこで筆者らは、この漢字と主題との関連性を「漢字の主題識別力」と考え、これを文献の自動分類に応用する実験を試みた^{11),12)}。ここでいう分類とは、与えられた文献を、既存の分類体系中の適切な主題分野に割り付けることである。

3.1 自動分類の方法

ある漢字 K_i が主題分野 C_j に対して持っている主題識別力を S_{ij} で表すと、 K_i は主題分野数 n だけの次元を持つベクトル

$$K_i = (S_{i1}, S_{i2}, \dots, S_{ij}, \dots, S_{in}) \quad \dots(6)$$

として表すことができる。 S_{ij} は、主題分野 C_j に対する K_i の主題識別力の強さに比例して大きい値を持つようにすればよい。そのため我々は、漢字の出現頻度傾向に基づいて S_{ij} を求めた。そして、ある文献が主題分野 C_j に対して持っている主題関連の強さを R_j とし、 R_j を S_{ij} を用いて以下のように表す。

$$R_j = \sum_i S_{ij} \quad \dots(7)$$

ただし、(7)式で加算の対象となるのは、その文献に出現する漢字の S_{ij} のみである。このようにして個々の文献について得られた、 n 個(主題分野数に相当する)の R_j のうち最も大きな R_j が得られた主題分野 C_j に、その文献を分類する。

漢字の主題識別力 S_{ij} としては、以下の式(8)-(11)により算出される漢字出現率、漢字出現偏差率 1、漢字出現偏差率 2、漢字補正出現率の四つを用いた。

$$\text{漢字出現率} = \frac{\text{主題分野 } C_j \text{ における漢字 } K_i \text{ の出現頻度 } (f_{ij})}{\text{漢字 } K_i \text{ の総出現頻度 } (\sum_j f_{ij})} \quad \dots(8)$$

$$\text{漢字出現偏差率 1} = \text{漢字出現率} - \text{漢字平均出現率} \quad \dots(9)$$

表6 分野出現偏差度2による一般漢字

	分野平均出現率	分野出現偏差度2		分野平均出現率	分野出現偏差度2
行	4.52120	7.07591	用	15.64373	8.10527
上	4.46362	2.18152	性	12.07345	7.07634
大	4.18689	0.79729	的	9.72401	6.05319
驗	4.17327	6.56703	方	9.01020	4.69633
算	4.15284	7.50045	化	8.32363	7.18648
特	4.14850	2.18339	分	8.10386	5.95121
休	3.91449	9.51323	合	7.60983	8.99242
検	3.85196	3.77922	形	6.09554	9.50205
記	3.85010	4.56226	特	6.07820	2.60904
場	3.82720	4.51588	実	6.07511	2.40952
論	3.58080	7.61199	示	5.90671	6.83614
効	3.53065	4.58107	結	5.90548	7.96020
中	3.48979	3.07059	線	5.86895	8.65057
本	3.41241	4.55321	変	5.84914	2.04360
人	3.41055	3.37391	間	5.77237	2.63309
生	3.38765	4.39121	成	5.74513	6.69121
現	3.37588	3.27832	対	5.57055	4.92322
開	3.36041	2.50383	解	5.53464	5.50820
率	3.32078	3.13039	果	5.08147	3.66423
応	3.30283	4.42518	要	5.00656	6.57591
表	3.26073	6.69333	述	4.99542	7.40383
明	3.24959	0.78571	高	4.80040	5.67786
基	3.22173	1.84328	作	4.74159	6.92462
案	3.15177	6.10283	出	4.58744	3.37648
使	3.04157	4.80655	構	4.57506	8.53018
点	3.00752	3.54619	一	4.55525	2.76630

ただし,

$$\text{漢字平均出現率} = \frac{\sum_i f_{ij}}{\sum_i \sum_j f_{ij}}$$

$$\text{漢字出現偏差率 } 2 = \frac{\text{漢字出現率} - \text{漢字平均出現率}}{\text{漢字平均出現率}} \quad \dots(10)$$

$$\text{漢字補正出現率} = \frac{\text{ある主題分野に対する補正後の出現頻度}}{\text{ある漢字の補正後の総出現頻度}} \quad \dots(11)$$

ただし,

$$\text{漢字補正後の出現頻度} = \frac{\text{ある出題分野における漢字の出現頻度}}{\text{分野比率}}$$

$$\text{分野比率} = \frac{\text{その主題分野に属しているサンプル・レコード}}{\text{全サンプル・レコード数}}$$

漢字の出現傾向をみると、ある主題分野における出現頻度が他の主題分野と比較して特に高いもの、出現頻度の多少にかかわらず、ある特定の主題分野に集中して出現するもの、どの主題分野にもまんべんなく出現するもの等がある。前述の(8)-(11)の四つの S_{ij} 算法は、これらの出現傾向を主題との関連性の観点から数量化するように試みたものである。主題分野における漢字の出現傾向を主題内容との関連で分析する場合には、漢字の出現頻度データを抽出したサンプル・レコード数が各主題分野間で異なることによって生じる影響を無視できない。そこで、前述した4番目の S_{ij} 算出

法ではその影響を除くよう考慮した。

漢字出現偏差率1は、ある主題分野に特に多く出現した漢字の S_{ij} を高くすることを意図している。それに対し漢字出現偏差率2は、ある主題分野に特に偏って出現する漢字の S_{ij} を高くすることを意図している。一方、漢字補正出現率は、サンプル・レコード数のバラツキ自体を補正しようとするものである。

文献の標題には、その文献の主題を表す重要語が用いられることが多い、長さがある程度限定されているため不要語が含まれる可能性が少ない等の特徴がある。そこで、我々の実験では標題中に出現する漢字を、主題識別力算出のためのデータの抽出源および分類実験の対象とした。

実験対象とした主題分野は、既述した12の分野であり、抽出した文献レコード数は21,960、総漢字数は251,661、異なり漢字数は1,404である。分類実験に使用したのは、各主題分野から45レコードずつの計540レコードである。この540レコードは、漢字の主題識別力算出のためのサンプル・レコード群とは重複しないように選択した。

3. 2 分類実験結果およびまとめ

この自動分類手法の評価は、再現率(自動分類の結果が、人手による分類結果をどの程度再現したかをその割合で示した尺度)を調べることによって行った。

まず、各算出法別の平均的な再現率を示したのが表7である。

漢字1文字が持つ主題識別力が分類のための情報としては無効であるならば、分類は不能となり、再現率は約8.3%(1/12)となるはずだが、実際にはこれを上回っている。したがって、漢字の主題識別力は、分類に有効な何らかの情報を与えていると考えることができる。

表8は、あらかじめ人手によってそれぞれの分野に分類されている文献が、自動分類の結果どの分野に分類されたかを示している。ここで使用された S_{ij} は、漢字補正出現率である。例えば、生体工学(カテゴリー番号4)では、正しく分類された

表7 Sij 算出法別再現率

主題識別力(S_{ij})	再現率
漢字出現率	23.5 %
漢字出現偏差率1	40.8 %
漢字出現偏差率2	46.7 %
漢字補正出現率	44.2 %

件数が31、誤って分類された件数が14である。なお‘NO’の列は分類不能を意味する。

主題分野ごとに求めた再現率は、分野によって大きな差異がみられる。例えば、電力工学やシステム工学一般では、93.3%、80.0%とそれぞれ高くなっているのに対し、計算機・ハードウェアや電子工学では0となっており、極めて対照的である。これは、この自動分類手法では、電力工学、システム工学一般に属する文献の分類は成功率が高いのに対し、計算機・ハードウェア、電子工学分野の文献の場合は低いことを示している。

「太陽光発電用構造物の設計における風荷重の考察」(カテゴリー10)は、成功例のひとつである。一方、「カラーファクシミリとカラー複写機」、「LEDアレーを用いた光学プリンタ」、「高性能集積回路の設計」はいずれもカテゴリー7に分類されるべきところが、カテゴリー8、2、11にそれぞれ誤って分類されている。

この分類法には、漢字1文字に着目しているため、入力時に分かち書き等をする必要がないことや、未知語に対しても漢字書きされていれば対応できる等の利点がある。

しかし、表8および上述した例から明らかなように、漢字の主題識別力のみでは、完全な分類をすることはできなかった。この原因としては、日本語の科学技術文献ではカタカナ書きされた語が多く使用されていることがあげられる。これらの語はいわゆるテクニカルターム(科学技術用語)で、キーワード的な働きをするものが多いが、今回の実験ではこれらを考慮しなかった。また、漢字には、1文字のみで用いられるもの(例えば、血、

表 8 自動分類実験結果

カテゴリー	再現率 (%)	カテゴリー番号														TOTAL
		1	2	3	4	5	6	7	8	9	10	11	12	NO		
1 計測工学 計測機器	33.3	15	3	6	4	3	2	0	0	1	5	0	6	0	45	
2 電磁気学 光学	48.8	2	22	6	1	5	0	0	1	1	5	1	1	0	45	
3 電子物性 磁性 光物性	66.6	1	6	30	0	2	0	0	0	3	3	0	0	0	45	
4 生体工学	68.8	1	4	2	31	1	2	0	2	0	1	1	0	0	45	
5 システム制御工学一般	80.0	2	0	0	3	36	3	0	0	0	1	0	0	0	45	
6 制御工学	46.6	2	1	0	2	10	21	0	0	1	8	0	0	0	45	
7 計算機方式 ハードウェア	0.0	5	5	2	1	5	4	0	8	3	5	1	5	1	45	
8 計算機利用技術	37.7	1	1	0	1	9	1	0	17	4	7	1	3	0	45	
9 電気工学一般	33.3	2	3	3	1	7	0	0	0	15	11	0	3	0	45	
10 電力工学	93.3	0	0	0	0	2	0	0	0	1	42	0	0	0	45	
11 電子工学	0.0	2	5	22	0	6	1	0	0	3	4	0	2	0	45	
12 通信工学	20.0	4	8	1	2	13	1	0	4	0	3	0	9	0	45	
計	44.2 %	37	58	72	46	99	35	0	32	32	95	4	29	1	540	

肉, 光), 熟語の構成要素となる場合が多いもの(例えば, 利, 電, 識), どの主題分野にも一様に出現するもの(例えば, 用, 性, 法)などがあるが, これらの漢字の性質をすべて同一として扱ったこと等もその原因と考えられる。

4. 漢字の出現頻度特性を利用したキーワード抽出¹³⁾

キーワード自動抽出の一つの方法として, 文献集合に対して既に分野のような区分が与えられているという前提があれば, 各分野間で出現頻度分布に偏りのない語は不要語として取り除き, 残った語をキーワードとする方法も考えられる。長尾らの実験はこの例にあたる¹⁴⁾。筆者らは, この考え方を漢字1字を対象に通用した実験を行った。

4.1 キーワード抽出法

各漢字の平均的分布からのずれを測るパラメータとして, 式のうえでは(5-2)式で示した分野出現偏差度2と同じになる χ^2 値を用いた。 χ^2 値を利用する理由は, 特定の分野に集中する識別力のある

漢字を抽出するのに, 分布型の適合度検定の考え方を導入するためである。

本来ならば, 総出現頻度の度数に応じて χ^2_0 (適当な有意水準の下での検定基準値)をきめるべきであるが, ここでは一律に1,000とし, すべての漢字について(5-2)式により χ^2 値を算出した。そしてこの値が1,000を越えるものをその分野の特徴を表す重要漢字とし, この漢字を含む単語をキーワードとして抽出した。

4.2 キーワード抽出の実験結果と分析

一般にキーワードの定義自身があいまいで, 何をもってキーワードとするかを決定することは難しい。本実験の場合, 各文献に付与されたキーワードは, シソーラスファイルから選ばれたデスクリプタであるので, これを文章中の語と比較しても意味がない。

重要漢字に基づくキーワードの妥当性を評価するには, 本来ならば, 被験者を無作為に抽出して, 文献から重要語を選択させ, これと比較すべきであるが, この場合, 被験者はこの対象とする分野(電気工学)の専門的知識をある程度周知している

〈例 1〉

標記ループの特性改善を目的として、補間機能を有する位相検出器とディジタル平滑器を用いた新しい構成法を提案。この方法を用いて、定常位相差の生じない二重同期ループの特徴を保有したまま、出力位相ジッタを大幅に低減できることを示した。位相制御信号の変動成分をディジタル的に平滑化する方法を提案。

〈例 1—重要漢字によって抽出されたキーワード〉

位相検出器, ディジタル平滑器, 補間機能, 定常位相差, 出力位相ジッタ, 大幅, 位相制御信号

〈“もれ”と考えられるキーワード〉 二重同期ループ

〈“ノイズ”と考えられるキーワード〉 大幅

〈例 2〉

BoakyeらとPaugらによって得られた継続伝送線路の時間域における実現条件が、反射係数に関する基本的性質から得られることを示した。また従来知られていなかった終端を短絡もしくは開設した継続伝送線路の段数と反射インパルス応答列との関係を明らかにした。継続伝送線路を合成する問題について記述。

〈例 2—重要漢字によって抽出されたキーワード〉

継続伝送線路, 反射係数, 関する, 段数, 反射インパルス応答列, 関係, 問題

〈“もれ”と考えられるキーワード〉 Boakye, Paug

〈“ノイズ”と考えられるキーワード〉 関する, 関係, 問題

図 3 重要漢字によるキーワード抽出の実験例

必要があると考えられるので、適切な被験者を得ることは難しい。

そこでやや恣意的ではあるが、筆者らが無作為に抽出した抄録を一通り読み、その中で重要語と考えられるものをキーワードとすることにした。この方法で抽出した重要漢字と文章中のキーワードとを対比させ、重要漢字とキーワードとの関係をとらえることにした。図3でアンダーラインが引かれている漢字は X^2 値が1,000以上になったものである。

重要漢字からキーワードを決定する時、1字でも重要漢字を含む語をすべてキーワードと考えると、あまりにキーワードの数が多くなり、有効性が少ないことがわかった。科学技術文献の場合、そのキーワードは、テクニカルタームであると考

えてよい。これらの語の多くは通常、いくつかの単語(名詞)が連続した複合名詞であることが多い。また、科学技術文献の場合、テクニカルタームの多くが英語等の訳語からなる場合があることから、カタカナ書きされた語は、キーワードと考えてさしつかえない。

したがって、重要漢字からキーワードを拾う際、比較的長い漢字列(複合語)や、カタカナ列となりあう名詞等が目安になる。図3で示す二つの例では、「特性改善」「目的」「提案」「用いた」「実現条件」「基本的性質」等の一般語と数えられるものが抽出されていないが、一方、明らかに一般語と考えられる語「大幅」などがキーワードとして抽出されている。したがって、ノイズが多くなることが予想される。これにどう対処するかが今後の

課題となろう。また、固有名詞等をキーワードとしたいときにも問題が生じる。この方法では、あらかじめキーワード抽出の対象とする日本語テキストが単語単位に分けられている必要がある。また、既存のデータベースでは、付与されているキーワードにふりがな付けされているものが多い。これらのことを考慮すると、本稿で提案した方法を有効利用するには、いわゆる「漢字カナ変換プログラム」と併用するのが最も効果的であると思われる。

5. お わ り に

この研究を通して、日本語の文書において使用されている個々の漢字が、ある程度文意を代表しているという見通しを得た。今後の研究の発展の方向としては、漢字表記のされていない語(特に、カタカナ書きの語)の取り扱いが、一つの課題となりうるであろう。カタカナ文字は表意文字ではないので何らかの意味を持つ最小の単位としては、欧米語でやられているような語単位をとらざるを得ない。このとき、語(今の場合カタカナ表記され

る語)と漢字とが持つ情報量は直感的には同じとはいえない。したがって、情報量の異なる2種の記号例から頻度計算を行って文献分類あるいはキーワード抽出に結びつけるモデルを考える必要があるが、それぞれの情報量をどうとらえるかが今後の実験課題となる。

また、漢字だけについてみても、同じ漢字が文脈により、異なる情報量を持つ場合が考えられる。例えば、「的」は「的を絞る」という用法と「効果的」という用法では明らかに異なる意味情報を持つと思われるが、ここではこれらを区別せずに対等に扱ってきた。漢字が使われる位置(構文的な位置や語構成上の位置)が自動的に解析できるならば精度をあげるよい手がかりとなる。この意味で漢字カナ変換は自動的な解析をするプログラムとして利用可能である。

したがって、これらの解析がどの程度、処理結果に寄与するかについて検討し、さらにこれらを考慮したモデルによって実験することが、より良いキーワード抽出および自動分類を行うために必要であろう。

参 照 文 献

- 1) Luhn, H. P. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development. 2(2)159-168(1958)
- 2) Maron, M. E. Automatic Indexing; An Experimental Inquiry. Journal of ACM. 8(3)404-417(1961)
- 3) Hamill, K. A. and Zamora, A. The Use of Titles for Automatic Document Classification. Journal of ASIS 31(6)396-402(1980)
- 4) Borko, H. and Bernick, M. Automatic Document Classification. Journal of ACM. 10(2)151-162(1963)
- 5) 西村, 岩坪. 計算機による文献の自動分類, 第6回情報科学技術研究会発表論文集. 107-120 (1968)
- 6) 西村, 岩坪. 邦文文献の自動分類, 第7回情報科学技術研究会発表論文集. 67-78(1969)
- 7) 中井. 機械補助索引(MAI)について[I]. 情報管理. 19(4)247-259(1976)
- 8) 中井. 機械補助索引(MAI)について[II]. 情報管理. 19(5)350-357(1976)
- 9) 細野, 後藤, 他. 漢字およびディスクリプターの頻度と主題分野の関連について. 東京サイエンティフィックセンターレポート(1984)
- 10) 後藤, 細野, 他. 出現頻度に基づく重要漢字と主題分野の関連. 第21回情報科学技術研究会発表論文集. 209-215(1984)
- 11) 細野, 守屋, 他. 漢字の出現頻度特性を用いた日

- 本語文献の自動分類, 日本アイビーエム JSI Research Report. TR 870006 (1984)
- 12) 梅田, 守屋, 細野, 他. 漢字の出現頻度情報を用いた日本語文献の自動分類. 情報処理学会自然言語処理研究会報告. 47(7)47-54 (1985)
- 13) 梅田, 諸橋, 他. 日本語文献に於ける漢字の出現頻度解析とその応用. 情報処理学会第 29 回全国大会 4 J-3, 1445-1446 (1984)
- 14) 長尾, 他. 日本語文献における重要語の自動抽出. 情報処理. 17(2)110-117 (1976)