

## 英文科学技術抄録文における名詞の決定

竹田 正幸<sup>†</sup> 須田 淳一郎<sup>††</sup>  
楠本 典孝<sup>†††</sup> 松尾 文碩<sup>†</sup>

自然言語処理における統語的曖昧さの発生は、主として単語が複数の品詞をもつことによる。統語解析の前に単語の品詞が決定できるならば、統語的曖昧さは激減する。英文科学技術抄録文では、名詞として現われる単語は延べ単語数の40%を占める。したがって精度の高い名詞決定法の開発は統語解析に寄与するところが大きい。本稿では、統語解析の前に、単語それ自身によって97%の確度で名詞が決定できることを示す。

### Identification of Nouns in Abstracts of Scientific and Technical Literature

MASAYUKI TAKEDA,<sup>†</sup> JUNICHIRO SUDA,<sup>††</sup> NORITAKA KUSUMOTO<sup>†††</sup>  
and FUMIHIRO MATSUO<sup>†</sup>

In natural language processing, syntactic ambiguity arises mainly owing to the words having multiple parts-of-speech. If the part-of-speech of each word in sentences is decided on before the syntactic analysis, the ambiguity will be decreased sharply. Since 40% of words appeared in the abstracts of scientific and technical literature are nouns, a method for identifying nouns with high accuracy would be helpful in the syntactic analysis. This paper shows that 97% of the nouns in sentences can be identified by each word itself, before the syntactic analysis.

#### 1. まえがき

自然言語文の統語解析における統語的曖昧さの発生は、自然言語処理研究の進展を阻む大きな障壁である。英文の場合、この曖昧さは主として多品詞単語が多いことに起因している。統語解析の前に単語の品詞が決定できるならば、統語的曖昧さは激減する。もちろん、すべての単語の品詞を統語的制約あるいは意味的制約なしに決定できるわけではない。しかし、文中で最も多数を占める名詞について単語それ自身によって品詞が決定できれば、統語解析の複雑さと統語構造の曖昧さが大きく減少する。このように統語情報や意味情報を使うことなく、単語それ自身で名詞を決定する方法を、ここでは名詞決定法と呼ぶことにする。本稿では、英文科学技術抄録文についての名詞決定法を提案する。

英文科学技術抄録文では、名詞として現れる単語は、延べ単語数の40%を占める(6.3節参照)ので、精度の高い名詞決定法の開発は統語解析に寄与するところが大きい。本稿で提案する名詞決定法は、単語が文中で名詞句の核となる名詞として現れる相対頻度がある閾値を超えていたとき、名詞と判定する。この相対頻度を算出するには、膨大なコーパスを必要とする。しかし、英文科学技術文はその分野の専門家にしか理解できないため、そのコーパスを人手で大量につくることはきわめて困難である。そこで、名詞句の核となる名詞の大部分がその後置語によって決定できることに着目して、中核名詞についての疑似コーパスを機械的に大量に作成し、それから近似的な相対頻度を求める方法をとった。この結果、97%の精度をもつ名詞決定法が得られることが判明した。

対象とした英文科学技術文は、INSPECテープ<sup>1)</sup>の抄録文である。INSPECテープは、英国IEE(the Institution of Electrical Engineers)が1969年から提供している代表的な英文二次文献データで、物理学や電気工学、電子工学、制御工学、計算機科学、情報工学の分野の文献の書誌的情報と抄録を含んでいる。

<sup>†</sup> 九州大学工学部

Faculty of Engineering, Kyushu University

<sup>††</sup> 九州大学大学院工学研究科修士課程

Master Course, Division of Engineering, Graduate School, Kyushu University

<sup>†††</sup> 松下電器産業(株)

Matsushita Electric Industrial Co., Ltd.

## 2. 被修飾名詞と修飾語

英語には、名詞を他の名詞の修飾語(noun modifier)として使う用法がある。本稿では、他の名詞を修飾する名詞を修飾名詞(modifier noun), 修飾される名詞を被修飾名詞(modificant noun)と呼ぶことにする。例えば, database systemにおいて database が修飾名詞で, system は被修飾名詞である。また, the document database では, database が被修飾名詞である。ここでは、名詞が修飾名詞として生起していないとき、その名詞は被修飾名詞として生起しているという。修飾・被修飾については連言句の場合など難しい問題があるが、これについては 3.2 節で例示する。

次の三つの例文で被修飾名詞の生起を太字で示した。また、斜体字の単語列は動詞句である。

**例文 1** The starburst **observations** are a major motivation for the consideration of this **model** since the extreme **conditions** are observed.

**例文 2** The properties of the black **hole** and the nonthermal **radiation** from its **environment** are calculated under the **assumption** that the mass **influx** is constant.

**例文 3** An approximate **solution** to these **equations** is determined using a Galerkin **technique** involving polynomial and trigonometric **functions**.

例文 1 では、被修飾名詞の後置語は動詞句と for, of, since である。例文 2 では、後置語は動詞句、of, and, from, that であり、例文 3 では、動詞句、to, involving, 文末のピリオドである。被修飾名詞の後置語の多くは、動詞句、現在分詞、過去分詞、of, for, from のような前置詞、and, since のような接続詞、文末ピリオドである。文末ピリオドも語とみなすことすれば、被修飾名詞は多くの場合、後置語によって決定できる。

## 3. 後置語による被修飾名詞の決定

基本的には、名詞の後置語が非名詞であればその名詞は被修飾名詞である。一般的には、文

This IR system saves researchers labor.

\* 1,487 の科学技術抄録文について、動詞の文型を調査したが、間接目的語と直接目的語、あるいは第 1 目的語と第 2 目的語が隣接する用法は 2 例だけであった。科学技術文では、目的語である名詞句が比較的長いことが多く、前置詞を使って二つの名詞句の関係を明示する傾向があるためであると考えられる。また、この 2 例とも、名詞句の先頭に冠詞を伴っており、被修飾名詞が隣接する用法ではなかった。なお、本稿の例文は、直前のものだけが例文としてつくったものであり、他の例文はすべて INSPEC テープの抄録文からとったものである。

のように、被修飾名詞 researchers のあとに名詞 labor が続く場合もあるが、このように目的語のあとに目的語が隣接する用法は、科学技術文ではまれである\*。したがって、後置語による被修飾名詞決定手続きは、基本的には、名詞に非名詞が後続する 2 単語列を決定する手続きということになる。

しかし、動詞の ing 形は現在分詞および名詞として出現する。例えば、動詞 engineer の現在分詞形 engineering は、「工学」を意味する場合には、通常、名詞に分類される。句

the software **engineering** of two important classes of computer systems

では、engineering は被修飾名詞として生起している。この場合、後置語 of で被修飾名詞であることが判定できる。一方、句

information systems engineering the design, implementation and evaluation of the human-machine interface

では、engineering は現在分詞であり、この場合は後置語 the で現在分詞であることがわかる。ところが、句

two injection moulded semicrystalline engineering thermoplastic materials

や

X 11 computer assisted software engineering integrated tool sets

の場合、engineering の後置語は形容詞あるいは過去分詞であるが、このことにより engineering が被修飾名詞と判断することはできない。したがって、ing 形の後置語が冠詞、形容詞、過去分詞の場合、ing 形を被修飾名詞とするわけにはいかない。形容詞と名詞の両方の品詞をもつ語についても、ing 形と同様な判断をしなければならない。

名詞の品詞をもつ語は、後置語が非名詞であれば被修飾名詞であるものと、後置語が冠詞、形容詞、過去分詞形のときは被修飾名詞としてはならないものに分けられる。そこで、被修飾名詞となりうる語を次の二つの範疇に分ける。

- A) 名詞として生起する可能性が高い語。
- B) ing 形のように名詞としても名詞以外の品詞としても生起する語。

B 以外の非名詞は、B の語の被修飾名詞決定のために次の二つの範疇に分ける。

- C) 冠詞、形容詞、過去分詞形。
- D) 前置詞、副詞、接続詞、動詞など、A, B, C 以外の非名詞および文末ピリオド。

被修飾名詞決定手続きは、まず文中の単語を A, B,

C, D に一意的に分類する。その後、A の単語の後置語が B, C, D の単語であるとき、A の単語を被修飾名詞とする。B の単語については、D の単語が後置語の場合、被修飾名詞と判定する。

### 3.1 語の分類

後置語によって、被修飾名詞を決定するためには、まず単語を A, B, C, D に分類しなければならない。このためには、単語の品詞を決める必要がある。動詞に関しては、著者らが開発した動詞決定法<sup>2)</sup>をもとに、連続単語列としての動詞句を 97% の精度で決定する方法を開発している。したがって、動詞句に関しては、生起時の品詞をほぼ決定することができる。

それ以外の品詞に関しては、あらかじめ各単語に 1 品詞を割り当てる。割当ての基本方針は、単語の原形が研究社英和中辞典第 5 版<sup>3)</sup>（以下、中辞典と略称する）の見出し語と一致した場合、その見出し語の最初の語義（第 1 語義）の品詞をその単語の品詞とする。日常文と科学技術文とでは、語義の重要度に差異があるので、この方法には問題がないわけではない。この点については 6 章で論じる。単語の原形が見出し語になかった場合は、語尾が ing と ed 以外の単語は名詞とした。科学技術抄録文における異なり語数は延べ単語数の平方根に比例することがわかっている<sup>4)</sup>。すなわち、科学技術文では新しい単語が出現する度合が日常文より高い。これらの単語は物質名、システム名などでほとんどが名詞である。名詞でない可能性は少ないながらもあると思われるが、INSPEC テープの抄録に生起する 20 万以上の単語を調査するのは困難な作業であるので、すべて名詞とした。

中辞典における小文字で始まる单一語 (single word) の見出し語は 38,910 語であり<sup>\*</sup>、変化形を補う

表 1 語の分類  
Table 1 Categories of words.

分類	中辞典にある語	中辞典にない語
A	第 1 語義が名詞の語	語尾が ing, ed でない語
B	第 2 語義以降に名詞をもつ語、 現在分詞形	語尾が ing の語
C	冠詞 (the, a, an), 第 1 語義が形容詞、 過去分詞形	語尾が ed の語
D	第 1 語義が前置詞、 by, like, minus, plus, till, 第 1 語義が接続詞、 第 1 語義が副詞、 第 1 語義が間投詞、 動詞 (分詞形を除く) 文末のピリオド	

\* この中辞典のまえがきには、総収録語数は約 7 万 5000 語と書かれているが、一般語の見出し語数は意外に少ない。

表 2 品詞内訳

括弧内の数字は名詞語義を併せもつ語の数を表す  
Table 2 Number of words for all parts-of-speech.

品詞	单品詞	多品詞	多品詞かつ
			第 1 語義
形容詞	6,255	4,210 (2,677)	1,498 (1,228)
冠詞	2	2 ( 0 )	1 ( 0 )
副詞	3,686	700 ( 316 )	159 ( 52 )
助動詞	12	19 ( 11 )	2 ( 1 )
接続詞	13	45 ( 16 )	9 ( 3 )
間投詞	82	123 ( 103 )	37 ( 32 )
名詞	17,903	6,525 ( — )	2,898 ( — )
代名詞	42	116 ( 73 )	16 ( 5 )
前置詞	25	86 ( 35 )	27 ( 10 )
動詞	2,700	3,769 ( 3,610 )	1,003 ( 958 )

表 3 第 1 語義により選出した単語

Table 3 Words selected by primary meaning.

(a) 第 1 語義が前置詞の語

\* は名詞語義を併せもつ語

(a) Prepositions in primary meaning.

about	besides	into	to
above*	between	neath	toward
across	betwixt	notwithstanding	towards
against	beyond	of	under
along	circa	off*	underneath*
amid	despite*	on*	
amidst	during	onto	until
among	ere	over	upon
amongst	ex*	per	versus
at	for	qua	via
below	from	thro	with
beneath	gainst	through	within*
beside	in*	thru	without*

(b) 第 1 語義が接続詞の語

\* は名詞語義を併せもつ語

(b) Conjunctions in primary meaning.

albeit	if*	tho	wherever
altho	lest	though	whether
although	nor	unless	while*
and	or	whenever	whilst
because	since	whereas	
but*	than	whereupon	

(c) 第 1 語義が副詞で名詞の品詞をもつ語

(c) Adverbs in primary meaning each also having noun as its part-of-speech.

alias	impromptu	outdoors	upstage
altogether	inward	outward	upstairs
aside	mezzo	overside	uptown
ay	midtown	pellmell	westward
aye	nay	piecemeal	when
downstage	northeastward	presto	where
downstairs	northward	southeastward	whereabouts
downtown	northwestward	southward	wherefore
eastward	now	southwestward	why
everywhere	nowadays	tandem	yea
forward	nowhere	today	yes
hereafter	once	tomorrow	yesterday
how	out	tonight	

と 77,740 語になる。A, B, C, D への分類結果を表 1 に示す。

表 2 に、品詞ごとの中辞典の語数を示す。括弧内の数は、名詞の品詞を併せもつ語の数である。第 1 語義が前置詞、接続詞の語、および第 1 語義が副詞で名詞の品詞をもつ語を表 3 に示す。科学技術論文においてこれらの語が名詞として生起することないと判断し、D に分類した。間投詞は科学技術論文に出現しないと考えられるが、同様に、第 1 語義が間投詞の語を D に加えた。

しかし、形容詞に関しては、第 1 語義が形容詞で名詞の品詞をもつ語は 1,228 語ある。この中には、single, simple など科学技術論文には名詞として生起しないと思われる語が含まれる一方、current, thermal のように、科学技術論文に形容詞および名詞の両品詞で頻出する語もある。これらの語については単語ごとの分類を行う必要があるが、語数が多いため、ここでは第 1 語義が形容詞の 7,753 語すべてを形容詞と考え、C に分類した。

代名詞については、第 1 語義が代名詞である 58 語に、that, its, these, this, those, he, it の 7 語を加え、主格・所有格・目的格ごとに表 4 のように分類した<sup>\*</sup>。動詞および助動詞については、動詞句決定法により文の動詞句がほぼ決定可能であるため、文から動詞句を除いた単語列を対象に名詞の決定を行うことにして、第 1 語義が動詞の語であっても、名詞の品詞をも

表 4 代名詞の分類  
Table 4 Classification of pronouns.

分類		語		
A	anybody	himself	ourself	they
	anyone	it	ourselves	thine
	anything	itself	she	thyself
	everybody	me	somebody	un
	everyone	myself	someone	us
	everything	nobody	something	we
	he	none	thee	you
	hers	nothing	theirs	yours
	herself	oneself	them	yourself
	him	ours	themselves	yourselves
B	her	his	idem	
C	its	our	these	those
	my	their	this	your
D	that	whom	whichever	whomsoever
	which	what	whichever	whosoever
	who	whatever	whoever	
	whose	whatsoever	whomever	

\* 語 idem は「同著者（の）；同語（の）；同書物[典拠]（の）」の意味をもつため、her などと同様、B に分類した。

てば B に分類した。

一方、被修飾名詞候補は、名詞の品詞をもつ語と動詞の ing 形とした。ただし、名詞の品詞をもつ語であっても、上で C や D に分類した語については、被修飾名詞候補としない。第 1 語義が名詞である 20,801 語を、常に名詞である可能性が高いと考え、A に分類した。第 2 語義以降に名詞をもつ語は、動詞の ing 形とともに B に分類した。

以上の分類では、第 1 語義を重視して単語を選出したが、例えば、語 by には二つの同形異義語があり、一方は前置詞と副詞、他方は名詞の品詞をもつ。この場合、上の分類では名詞の品詞をもつ語として B に分類される。しかし、科学技術文では、by は名詞としてではなく前置詞として生起すると考えられる。そこで、このような語として、by, like, minus, plus, till の 5 語を前置詞として D に追加した。

### 3.2 決定手続きと評価

文中の単語を A, B, C, D に分類したあと、被修飾名詞を決定する手続きは前に述べた。すなわち、A の単語が被修飾名詞であるのは、B, C, D の単語が後置されたときであり、B の単語が被修飾名詞であるのは D の単語が後置されたときである。

この判定を簡単に実行するために、単語の被修飾指数(modificant index)を導入する。語  $w$  の被修飾指数を  $m(w)$  で表し、文が

$$w_1 w_2 \cdots w_n w_{n+1}$$

ここで、 $w_{n+1}$  は文末ピリオド。

であるとすると、 $m(w_i) - m(w_{i+1}) \geq t$  ( $i=1, \dots, n$ ) ならば、 $w_i$  を被修飾名詞とする手続きを考える。いま、A, B, C, D に属する単語の被修飾指数を、それぞれ、4, 2, 1, 0 とすると、 $t=2$  のとき、上記の判断基準を満たした手続きになる。すなわち、後置語との被修飾指数の差が 2 以上のとき、被修飾名詞であると判定する。

2 章の例文の各単語に被修飾指数を付与した。[と]で囲まれた数字が各単語のもつ被修飾指数を表わしている。

例文 1 The<sub>[1]</sub> starburst<sub>[4]</sub> **observations**<sub>[4]</sub> are<sub>[0]</sub> a<sub>[1]</sub> major<sub>[1]</sub> **motivation**<sub>[4]</sub> for<sub>[0]</sub> the<sub>[1]</sub> **consideration**<sub>[4]</sub> of<sub>[0]</sub> this<sub>[1]</sub> **model**<sub>[4]</sub> since<sub>[0]</sub> the<sub>[1]</sub> extreme<sub>[1]</sub> **conditions**<sub>[2]</sub> are<sub>[0]</sub> observed<sub>[0]·[0]</sub>

例文 2 The<sub>[1]</sub> **properties**<sub>[4]</sub> of<sub>[0]</sub> the<sub>[1]</sub> black<sub>[1]</sub> **hole**<sub>[4]</sub> and<sub>[0]</sub> the<sub>[1]</sub> nonthermal<sub>[4]</sub> **radiation**<sub>[4]</sub> from<sub>[0]</sub> its<sub>[1]</sub> **environment**<sub>[4]</sub> are<sub>[0]</sub> calculated<sub>[0]</sub> under<sub>[0]</sub> the<sub>[1]</sub> assumption<sub>[4]</sub> that<sub>[0]</sub> the<sub>[1]</sub> mass<sub>[4]</sub> **influx**<sub>[4]</sub> is<sub>[0]</sub> constant<sub>[1]·[0]</sub>

例文 3 An<sub>[1]</sub> approximate<sub>[1]</sub> solution<sub>[4]</sub> to<sub>[0]</sub> these<sub>[1]</sub> equations<sub>[4]</sub> is<sub>[0]</sub> determined<sub>[0]</sub> using<sub>[2]</sub> a<sub>[1]</sub> Galerkin<sub>[4]</sub> technique<sub>[4]</sub> involving<sub>[2]</sub> polynomial<sub>[1]</sub> and<sub>[0]</sub> trigonometric<sub>[1]</sub> functions<sub>[2]·[0]</sub>

この三つの例文では、被修飾名詞がこの手続きで正しく判定されることがわかる。

1989年配布のINSPECテープから無作為に抽出した350抄録の1,487文に対し、動詞句と被修飾名詞についてのコーパスを入手により作成し、この決定法を適用したところ、9,250語を被修飾名詞と判定した。このうち正しく判定したのは7,861語であった。このコーパスは、8,182語の被修飾名詞を含むので、情報検索の用語を借りるならば再現率(recall rate)は96.1%，精度(precision)は85.0%ということになる。

被修飾名詞でない語を被修飾名詞とする誤りのうち1割程度は、as well as, in order to, in terms of, a number of, with respect to, by means of等の成句中のA, Bの語を被修飾名詞と判定したものであった。これらの誤りは、文中の成句を決定できれば除くことができる。これ以外の誤りのうち、主なものは次の二つである。一つは、連言句と選言句の問題である。例えば、句

the<sub>[1]</sub> research<sub>[4]</sub> and<sub>[0]</sub> development<sub>[4]</sub> program<sub>[4]</sub>において、researchは、developmentとともにprogramを修飾する修飾名詞であるが、これを被修飾名詞と誤認してしまう。後置語によって被修飾名詞を判定する方法は、連言句と選言句に対しては正しい結果をえたえない。

もう一つの誤りは、例えば、句

A<sub>[1]</sub> low<sub>[1]</sub> power<sub>[4]</sub> single<sub>[1]</sub> chip<sub>[4]</sub> speech<sub>[4]</sub> processor<sub>[4]</sub>

では、修飾名詞powerを被修飾名詞と誤認している。この場合、A low powerとsingle chip speech processorという二つの名詞句が隣接して現れている。二つの名詞句が、前置詞などで結合するのではなく、隣接して現れる場合、間接目的語と直接目的語が隣接している可能性もあるが、大部分は、修飾名詞と形容詞が修飾語として混在している場合であると考えられる。次章で述べる疑似コーパスの作成においては、このようなpowerを被修飾名詞としないように手続きを変更した。変更した手続きは、上記のコーパスに対して、8,081語を被修飾名詞と判定し、このうち正しく判定したのは7,508語であった。すなわち、再現率は91.8%に低下したが精度は92.9%に向上了。

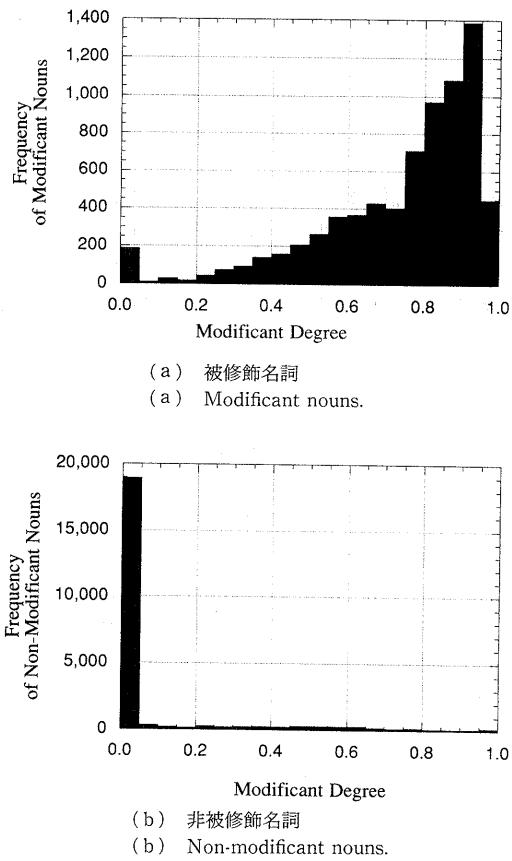


Fig. 1 被修飾度と生起頻度  
(a) 被修飾名詞  
(a) Modificant nouns.  
(b) 非被修飾名詞  
(b) Non-modificant nouns.

図1 被修飾度と生起頻度

Fig. 1 Relationship between modificant degree and frequency.

#### 4. 疑似コーパスからの被修飾度の算出

単語が被修飾名詞として生起する相対頻度(確率)を被修飾度(modificant degree)と呼ぶことにする。被修飾度を算出するためには、被修飾名詞に関する大量のコーパスが必要である。しかし、この作成は多大な労力を要する。そこで、INSPECテープ抄録文924,720文を対象に3章の被修飾名詞決定手続きによって疑似コーパスを作成し、この疑似コーパスから近似的に被修飾度を算出した。

3.2節で述べた人手で作成したコーパス中の被修飾名詞延べ8,182語、および被修飾名詞でない語延べ23,870語について、被修飾度に関する度数分布を、それぞれ、図1(a), (b)に示す。ただし、英大文字、数字、特殊記号を含む語については、被修飾度を1とした。

被修飾度0.1付近を境に、被修飾名詞とそうでない語とに分かれる。また、被修飾名詞の頻度が、被修飾度0のところで高くなっているが、この主な理由は、

第1語義を形容詞とする語が被修飾名詞として生じたためである。

## 5. 被修飾度の閾値と決定成功率

単語の被修飾度が大きいということは、その単語が名詞として生じる確率が高いと考えられる。名詞と形容詞の語義をもつ単語の場合、修飾語として生じた場合は形容詞語義の確率が高いという場合もあるが、ここでは単純に、このような語は名詞とみなす。

そこで、本稿で提案する名詞決定法では、被修飾度0.1付近に閾値を設定し、その閾値以上の被修飾度をもつ語を名詞とする。この名詞決定に先立ち、動詞句は別の方法で決定し、それらの語については、名詞であるかどうかの対象とはしない。以下の文では、[と]で囲まれた数が被修飾度を表わす。動詞句は、名詞候補の対象外であるため、被修飾度をあたえない。

**例文1** The<sub>[0]</sub> starburst<sub>[0.1860]</sub> **observations**<sub>[0.9408]</sub> are a<sub>[0]</sub> major<sub>[0]</sub> **motivation**<sub>[0.9690]</sub> for<sub>[0]</sub> the<sub>[0]</sub> **consideration**<sub>[0.9448]</sub> of<sub>[0]</sub> this<sub>[0]</sub> **model**<sub>[0.8147]</sub> since<sub>[0]</sub> the<sub>[0]</sub> extreme<sub>[0]</sub> **conditions**<sub>[0.8785]</sub> are observed.

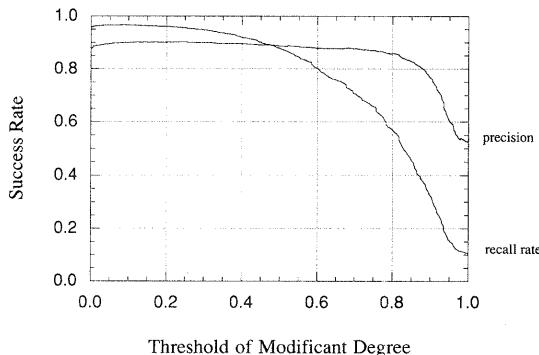


図2 閾値と成功率

Fig. 2 Relationship between success rate and threshold.

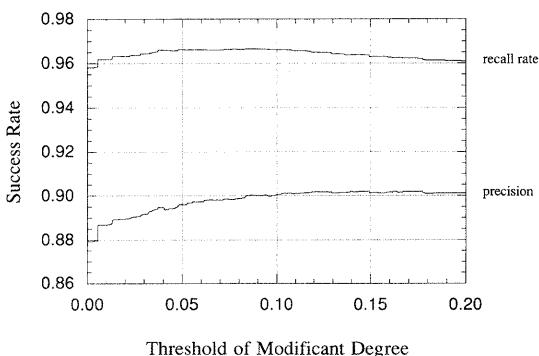


図3 閾値0.1付近の成功率の推移

Fig. 3 Success rate-threshold relationship near 0.1.

**例文2** The<sub>[0]</sub> **properties**<sub>[0.9647]</sub> of<sub>[0]</sub> the<sub>[0]</sub> black<sub>[0]</sub> hole<sub>[0.3755]</sub> and<sub>[0]</sub> the<sub>[0]</sub> nonthermal<sub>[0.0827]</sub> **radiation**<sub>[0.5300]</sub> from<sub>[0]</sub> its<sub>[0]</sub> **environment**<sub>[0.9129]</sub> are calculated under<sub>[0]</sub> the<sub>[0]</sub> **assumption**<sub>[0.9591]</sub> that<sub>[0]</sub> the<sub>[0]</sub> mass<sub>[0.3549]</sub> **influx**<sub>[0.7857]</sub> is constant<sub>[0]</sub>.

**例文3** An<sub>[0]</sub> approximate<sub>[0]</sub> **solution**<sub>[0.8831]</sub> to<sub>[0]</sub> these<sub>[0]</sub> **equations**<sub>[0.9344]</sub> is determined using<sub>[0.0054]</sub> a<sub>[0]</sub> Galerkin<sub>[1]</sub> **technique**<sub>[0.9287]</sub> involving<sub>[0.0130]</sub> polynomial<sub>[0]</sub> and<sub>[0]</sub> trigonometric<sub>[0]</sub> **functions**<sub>[0.8577]</sub>.

上記例文では、被修飾度が0.1以上の単語はすべて名詞であり、0.1未満のものは、すべて非名詞であることがわかる。いま単純に、名詞と判定した語が連続している場合にはその最終語を被修飾名詞とすると、被修飾名詞の決定に成功している。もちろん、この方法では、3.2節で述べたような、修飾名詞が被修飾度の小さい後置形容詞によって被修飾名詞であると誤認される可能性やing形が被修飾名詞と誤認される可能性は残っている。いずれにしても、名詞句の統語構造の問題は本稿では扱わない。

この決定法で、閾値を変化させたときの再現率・精度の推移を、図2に示す。図1から被修飾度が0.1の付近に名詞と非名詞を分ける点があるようにみえる。そこで、この付近の再現率と精度を拡大して示したのが図3である。図3に示すように、再現率は、閾値0.08から0.09付近で最大となり、この範囲において精度が最大となるのは、閾値が0.086付近のときである。閾値が0.086の場合の再現率は96.65%であり、精度は90.03%であった。

## 6. 考察

### 6.1 改善点

本方式は、高い決定精度をもつが、改善の余地が残されている。3.1節で述べたように、名詞の品詞をもつ語でも、CやDに分類されることがある。CとDの単語は疑似コーパス作成において、被修飾名詞にはならないため、これらの単語の被修飾度は0である。これらのうち、第1語義が前置詞、接続詞などの語については、数が少ないため、人手により、科学技術論文では名詞としての生起がないと判断することができた。しかし、第1語義が形容詞で名詞の品詞をもつ1,228語については、各語について吟味することができなかった。この理由で、名詞と判定することに失敗した単語としては、potential, component, constant, current, variant, radical, perpendicular, complex, characteristic, alternative, tutorial, level, integralなどがある。これらは、本来Bに分類すべき語であった。中辞

典では、これらの語の大半が名詞の語彙の中に「理」や「数」などの分野を示す情報を含んでいる。この情報を利用して、これらの語を B に分類することもできる。

中辞典には、名詞の品詞をもたない見出し語が 14,482 語あるが、科学技術文では名詞として出現するものがある。それらの被修飾度は、本方式ではやはり 0 になる。この理由で名詞の判定に失敗した語としては、transform, quench などがある。この二つの語は中辞典では動詞の語義しかもたない。

以上の二つの理由に基づいて、被修飾度が 0 になっている単語を B に分類するためには、被修飾指数の条件を緩め、極端な場合、of の前置語をとり、名詞として出現する可能性のある語を科学技術文から抜き出して検討する必要がある。

## 6.2 多品詞名詞

被修飾度は、名詞でない語を見分ける能力があるが、多品詞単語について被修飾度に基づいて名詞であると判定することには問題がある。例えば、current は前に述べた理由により、本来ならば名詞と判定される。the leak current や the current amplifier の場合には、current は名詞であるが、current topics の場合は形容詞である。また、current system が「現システム」であるのか「電流系統」であるのかは、文脈に依存する。

結局、current のような語の品詞決定は意味にかか

わる問題であるので、本稿のように統語解析以前の問題を扱う段階においては、この品詞決定の問題を断念せざるをえない。しかし、統語解析の段階では、このような単語に関して被修飾名詞か名詞の修飾語であるかが判別できればよい。

名詞の修飾語には、形容詞と名詞がある。中辞典には、少なくとも名詞と形容詞の両品詞をもつ見出し語が 2,677 あるが、今回用いた 924,720 文の疑似コーパス中に生じた語は 1,788 語であった。このうち、993 語の被修飾度が 0 であり、被修飾度が 0.086 より小さいものは 1,034 語であった。このような語は、修飾語

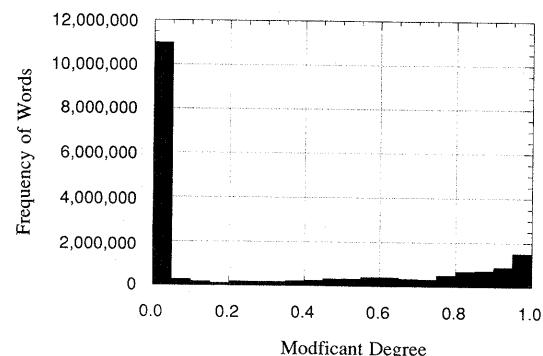


図 4 疑似コーパスにおける語の被修飾度と生起頻度

Fig. 4 Modificant degree—frequency relationship on quasi-corpora.

表 5 被修飾度 0.086 以上の高頻度語

Table 5 High frequent words with modificant degree  $\geq 0.086$ .

語	生起頻度	被修飾度	語	生起頻度	被修飾度
system	65693	0.7571	function	17821	0.8578
it	58769	0.8175	properties	17507	0.9648
results	45243	0.7825	problem	17216	0.9043
data	40555	0.5848	measurements	17155	0.9449
model	40185	0.8142	information	17140	0.6218
authors	39627	0.6945	theory	16145	0.9348
method	38536	0.9349	degrees	16138	0.2551
systems	36370	0.8566	number	16074	0.9048
temperature	30963	0.5526	effect	15885	0.9431
field	26461	0.6072	order	15774	0.7999
energy	26113	0.4511	study	15737	0.8242
time	24847	0.6284	parameters	15702	0.9056
analysis	24759	0.8735	conditions	15316	0.8786
they	24331	0.6037	state	15236	0.6764
surface	23884	0.4606	performance	15193	0.7396
structure	23833	0.8642	computer	14920	0.2535
use	23110	0.8007	methods	14824	0.9399
control	22309	0.4606	effects	14469	0.8945
design	22226	0.5840	development	13888	0.8108
phase	20707	0.4201	technique	13772	0.9288
power	20157	0.3209	frequency	13717	0.4435
process	19522	0.7047	distribution	13628	0.7730
author	19392	0.6762	show	13165	0.6137
range	18851	0.7184	paper	13164	0.7763
well	18007	0.5199	rate	13018	0.7205

表 6 被修飾度 0.086 未満の高頻度語

Table 6 High frequent words with modificant degree &lt; 0.086.

語	生起頻度	被修飾度	語	生起頻度	被修飾度
the	1601860	0.0000	these	47139	0.0000
of	959402	0.0000	were	46123	0.0000
and	557763	0.0000	using	45696	0.0054
a	469394	0.0000	used	41481	0.0000
in	392318	0.0000	or	37283	0.0000
to	380497	0.0000	between	34405	0.0000
is	319919	0.0000	two	34271	0.0000
for	253302	0.0000	also	33946	0.0000
are	193294	0.0000	presented	27658	0.0000
with	179535	0.0000	high	25034	0.0000
by	143544	0.0000	obtained	24777	0.0000
on	129887	0.0000	one	24693	0.0000
that	118623	0.0000	such	24645	0.0000
an	107777	0.0000	new	24596	0.0000
as	100851	0.0000	not	24424	0.0000
be	90066	0.0000	than	24274	0.0000
at	89445	0.0000	both	24150	0.0000
this	86618	0.0000	discussed	23935	0.0000
from	83890	0.0000	its	23107	0.0000
which	70761	0.0000	their	22593	0.0000
been	65042	0.0000	based	22565	0.0000
has	59735	0.0000	found	22291	0.0000
have	57242	0.0000	into	22161	0.0000
was	55243	0.0000	different	21344	0.0000
can	50693	0.0591	some	21203	0.0000

と考えてよい。また、被修飾度が 0 である 993 語のうち、複数形の生起がない語が 581 語あった。このような語は形容詞あるいは非可算名詞であると考えられる。いずれにしても、current のような単語はそれほど多くないため、人手で意味辞書をつくることが可能である。

### 6.3 被修飾度と単語の生起

被修飾度と単語の生起について、二、三調査した。まず、疑似コーパス文の延べ単語数は 18,892,543 で、このうち被修飾度が 0.086 未満の語は 11,218,837、0.086 以上の語は 7,673,706 で、名詞と判定したものは 40.6% であった。まえがきに示した数値は、これに基づいている。図 4 に、疑似コーパスに生起した全延べ単語についての被修飾度と頻度の関係を示す。

表 5 と表 6 に 0.086 以上の単語の高頻度語と 0.086 未満の単語の高頻度語を示す。表 5 と表 6 の高頻度単語に関しては、閾値 0.086 は名詞と非名詞をよく弁別しているようにみえる。

表 5 から、被修飾度 0.086 以上の高頻度語は、system, model, temperature などの専門語とともに author, paper など抄録文特有の語が多いことがわかる。一方、表 6 に示した被修飾度 0.086 未満の高頻度語としては、冠詞、前置詞、接続詞、動詞などがくなっている。

表 7 被修飾度 0.086 以上の語の最終音節

Table 7 Last syllables with modificant degree  $\geq 0.086$ .

最終音節	単語数	生起頻度	被修飾度	最終音節	単語数	生起頻度	被修飾度
tion	868	538346	0.7299	age	53	33734	0.4640
ing	2569	476380	0.2698	tures	43	32226	0.8497
tions	512	183374	0.9076	face	5	31960	0.5178
er	531	176498	0.2927	cy	94	31724	0.6543
ty	387	133973	0.7256	ple	22	30971	0.3904
ment	192	91961	0.7565	nal	56	28820	0.1553
y	256	87106	0.2187	ence	27	27928	0.7667
ture	53	86480	0.6718	um	90	26908	0.4331
tem	4	65882	0.7570	ters	90	26267	0.8740
sion	111	65332	0.5766	cess	5	25610	0.6790
el	22	58256	0.5916	ance	40	24032	0.7222
ments	114	57651	0.8992	sign	3	22397	0.5796
ry	142	50408	0.4800	trol	2	22338	0.4605
ta	22	49780	0.5270	tors	139	21797	0.8545
ter	120	48510	0.4101	fect	6	20272	0.7654
ties	180	47393	0.9358	els	13	20102	0.8537
sults	3	45247	0.7825	thor	1	19392	0.6762
gy	71	44326	0.6223	ics	62	19144	0.7803
od	2	42737	0.9150	per	32	18931	0.5980
sis	35	40035	0.8760	fects	4	18390	0.8506
thors	1	39627	0.6945	a	61	17730	0.6363
tor	153	38076	0.6113	ma	22	17324	0.3579
tems	2	36982	0.8575	lem	1	17216	0.9043
der	34	36400	0.4183	sure	16	17174	0.5097
ers	341	35533	0.7791	tics	24	17169	0.9322

表8 被修飾度 0.086 未満の語の最終音節  
Table 8 Last syllables with modificant degree < 0.086.

最終音節	単語数	生起頻度	被修飾度	最終音節	単語数	生起頻度	被修飾度
ly	1017	196775	0.0110	scribed	7	19569	0.0000
al	252	123336	0.0878	oped	3	18798	0.0000
cal	176	121837	0.0001	posed	15	18357	0.0000
ic	256	97578	0.0176	ied	11	18335	0.0000
ble	413	81425	0.0417	ar	6	16775	0.0016
tive	201	57838	0.0341	rent	5	16447	0.0001
ed	153	54296	0.0000	pared	3	15795	0.0000
tal	56	47506	0.0645	ate	80	15579	0.0193
lar	66	44114	0.0060	ered	25	15286	0.0000
tial	31	35282	0.0000	ized	79	15267	0.0000
tween	1	34405	0.0000	sured	7	15116	0.0000
so	6	33979	0.0005	ver	6	14793	0.0233
ous	92	32847	0.0000	gated	15	14466	0.0000
ent	37	31393	0.0436	id	15	14463	0.0014
sented	3	29244	0.0000	fied	51	14137	0.0000
en	61	27672	0.0029	lated	40	13502	0.0000
tained	10	27301	0.0000	bout	1	12998	0.0000
cussed	2	24053	0.0000	tain	16	12836	0.0037
to	17	24025	0.0015	put	2	11393	0.0000
tic	90	23983	0.0150	mined	3	11321	0.0000
mal	20	23916	0.0023	plied	6	11308	0.0000
tron	7	22573	0.0756	sive	69	11276	0.0003
ated	65	20726	0.0000	cial	17	10986	0.0000
served	7	20667	0.0000	tant	20	10863	0.0134
duced	11	20189	0.0000	tral	10	10541	0.0000

また、著者らは、辞書にない語の品詞を語尾によって決定する目的で、辞書の語の最終音節とその品詞との関係を調査したことがあるが、満足な結果を得られずにいた。そこで、今回算出した被修飾度を用いて、語の最終音節と名詞らしさとの関係をみるために、被修飾度 0.086 以上の語、0.086 未満の語について、最終音節ごとにまとめて被修飾度を算出した。高頻度の最終音節の被修飾度を、それぞれ表7と表8に示す。ここで、単語の最終音節は中辞典の分綴指示に従い、中辞典に記載されていない語については対象外とした。

表7では、tion, sion, ty, mentなどの名詞をつくる接尾辞およびそれらの複数形語尾の被修飾度の値が高いことがわかる。一方、表8では、副詞語尾 ly や形容詞語尾 al, cal, ble, tiveなどのほか、過去分詞形の語尾が多数出現している。これら最終音節ごとの被修飾度を利用して、疑似コーパス中に生起しない語についても、名詞か非名詞かの判定を行う方法が考えられる。ただし、中辞典にない語の最終音節を決定しなければならないので、分綴に関する研究を行う必要がある。

## 7. む す び

本稿で提案した名詞決定法は、英文科学技術抄録文に対して、97%の確度で名詞を決定できる。名詞と決

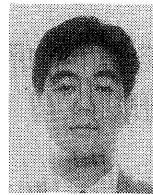
定された語の中には形容詞も混じっている。しかし、この決定法は統語解析の前段階で利用するものであるから、形容詞を名詞と誤認しても不都合はない。形容詞か名詞かの判別は、最終的には意味解析によらなければならない。

この決定法の成果を基に名詞句の決定法を研究中である。名詞句の決定において困難なことは、古くから指摘されているように、and/or によって結合された連言句/選言句と分詞形を伴った名詞句の統語構造を決定する問題である。前者については、この名詞決定法を使用することにより、71%の連言句/選言句の構造が決定できる<sup>5)</sup>が、この程度の精度では不十分である。この精度を上げ、後者の分詞形の問題を解決するためには、基本的な単語についての意味的分類が必要である。これまでの研究で英文科学技術抄録文の場合は、そのような基本的な単語は多くはないことがわかっている。

## 参 考 文 献

- 1) Aitchison, T. M., Martin, M. D. and Smith, J. R.: Developments Towards a Computer-Based Information Service in Physics, Electrotechnology, and Control, *Inform. Stor. Retr.*, Vol. 4, No. 2, pp. 177-186 (1968).

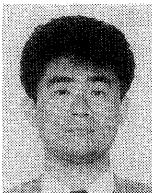
- 2) 竹田正幸, 松尾文碩: 英文科学技術抄録文における動詞の決定, 情報処理学会論文誌, Vol. 34, No. 9, pp. 1931-1936 (1993).
- 3) 小稻義男ほか編: 新英和中辞典, 第5版, 研究社 (1985).
- 4) Matsuo, F.: On Word Occurrence in Scientific and Technological Texts, 情報処理学会自然言語処理研究会資料, 46-2 (1984).
- 5) 日吉樹, 須田淳一郎, 竹田正幸, 松尾文碩: 英文科学技術文における連言問題, 第47回電気関係学会九州支部連合大会講演論文集, p. 688 (1994).  
 (平成6年11月28日受付)  
 (平成7年4月14日受付)

**須田淳一郎 (正会員)**

1970年生。1993年九州大学工学部電気工学科卒業。1995年九州大学大学院研究科電気工学専攻修士課程修了。同年(株)東芝入社。

**楠本 典孝**

1969年生。1992年九州大学工学部電気工学科卒業。1994年九州大学大学院工学研究科電気工学科専攻修士課程修了。同年松下電器産業(株)入社。

**竹田 正幸 (正会員)**

1964年生。1989年九州大学大学院総合理工学研究科情報システム学専攻修士課程修了。同年九州大学工学部電気工学科助手。自然言語理解、情報検索、パターン照合アルゴリズムに興味をもつ。日本ソフトウェア科学会会員。

**松尾 文碩 (正会員)**

昭和16年生。昭和41年九州大学大学院工学研究科電子工学専攻修士課程修了。工学博士。九州大学工学部電気工学科勤務。推論機構、自然言語理解、データベース、情報検索システム、エキスパートシステムの研究に従事。