英文科学技術文における単純名詞句の範囲決定

2L－5

丸木 健次† 柴田 誠† 日昔 吉樹‡ 竹田 正幸† 松尾 文碩†
†九州大学大学院システム情報科学研究科 ‡NTT

1. まえがき

英文科学技術抄録文を論理式へ変換する第一段階として，原子論理式の項に名詞句をそのまま単語列としてあてる方式が考えられる[1]．その名詞句の決定は次のような手順で行う．

(1) 被修飾名詞[2]の決定．

(2) 一つの被修飾名詞とその前方修飾語からなる単純名詞句の範囲決定．

(3) 単純名詞句をもとにした名詞句の範囲決定．

単純名詞句の範囲決定のためには先頭の語と末尾の語を決定しなければならない．被修飾名詞については名詞決定法[2]により98%の確度で決定できるので，前方修飾語の決定が問題となる．

2. 前方修飾語の決定

前方修飾語の決定については，辞書の品詞情報を用いる方法が考えられる．しかし，品詞の曖昧さ，単語の89%が辞書にないこと（表１）を考えると品詞情報はその決定には有効ではない．

そこで，1984年から1993年の10年分のINSPECテープ2,408,118文献の抄録文10,482,511文を調査し，統計的手法により決定することを試みた．

冠詞 the の次の語は前方修飾語か被修飾名詞と考えられる．これより，the の後に語 w が生起する頻度を $f_{\text{the}}(w)$ としたときある閾値 t を定め，

$$f_{\text{the}}(w) > t$$

となる語を前方修飾語とする方法が考えられる．

表１ 辞書にない単語数

$f(w)$			単語数	辞書にない単語数
1	～	9	346131	329256
10	～	99	64432	51496
100	～	999	20342	11409
1000	～	9999	7304	2064
10000	～	99999	2412	225
100000	～		276	23

表２ 高頻度語の生起

$f_{\text{the}}(w)$	$f(w)$	$f_{\text{the}}(w)/f(w)$	w
844	19470526	0.000043	the
127	11470994	0.000011	of
288	6863571	0.000042	and
12901	5819398	0.002217	a
8008	4798647	0.001669	in
323	4616776	0.000070	to
646	4071470	0.000159	is
39	3009524	0.000013	for
80	2402460	0.000033	are
14	2159270	0.000006	with
147	1750508	0.000084	by
3398	1572760	0.002161	on
14	1571699	0.000009	that
203	1299853	0.000156	an
6656	1262086	0.005274	as
991	1178265	0.000841	be
23	1072053	0.000021	this
965	1071746	0.000900	at
33	1013825	0.000033	from
12	979632	0.000012	which

しかし，高頻度の前置詞の中には $f_{\text{the}}(\text{'at'}) = 965$, $f_{\text{the}}(\text{'in'}) = 8{,}008$ のように，the の後置語としての生起頻度が比較的高いものがある（表２）．一方，このような前置詞が前方修飾語に含まれないように閾値

Decision on Simple Noun Phrase in Scientific and Technical Documents
Kenji Maruki†, Makoto Shibata†, Yoshiki Himukashi‡, Masayuki Takeda† and Fumihiro Matsuo†
† Graduate School of Infomation Science and Electorical Engineering, Kyushu University, Hakozaki, Fukuoka 812-81, Japan
‡ NTT

t を定めると形容詞 outermost $f_{\text{the}}(\text{'outermost'}) = 788, f(\text{'outermost'}) = 966$ も前方修飾語に含まれなくなる．このように $f_{\text{the}}(w) > t$ とする方法では決定できない．

そこで，比 $f_{\text{the}}(w)/f(w)$ を考え，この比がある閾値 t に対して

$$f_{\text{the}}(w)/f(w) > t$$

となる語を前方修飾語とする方法を考える．

このさい，低頻度語，例えば $f(w) = 1$ である語は単語の41%を占めている．このような低頻度語は前方修飾語である可能性が高いと思われるので，これらは判定の対象とはしない．

ここでは，この閾値を次のようにして求めた．形容詞の品詞のみをもつ語 4,590 語と前置詞の品詞のみをもつ語 22 語に対して，前者を前方修飾語，後者を前方修飾語にならない語としたい．この二つを分ける点を閾値とする．具体的には，形容詞の累積相対頻度 C_{adj} と前置詞の累積相対頻度 C_{prep} を求め，評価関数 $(1-C_{adj})*C_{prep}$ を最大とする点を閾値とする（図1）．

この方法で求めた閾値 $t = 0.0119$ を用いると，$f_{\text{the}}(w)/f(w) > t$ となる語は 161,181 語，$f_{\text{the}}(w)/f(w) \leq t$ となる語は 279,717 語となる（表3）．このとき，前方修飾語ではない語のなかには前置詞の品詞のみをもつ語 22 語のうち 18 語，形容詞の品詞のみをもつ語 4590 語のうち 1,433 語が含まれている（表4）．

ここで，前述のように低頻度語を前方修飾語として除くと表3のように $f_{\text{the}}(w)/f(w) \leq t$ となる語の割合は減少していく．

このとき，$f(w)$ がどのくらいから低頻度になるかという問題があるが，表4の前方修飾語ではない語のなかの前置詞の品詞のみをもつ語，形容詞の品詞のみをもつ語の頻度について閾値を求めたときの評価関数を用いると $f(w) \geq 100$ のとき最大となる．

3. むすび

この論文では，単純名詞句の範囲決定に必要な前方修飾語の決定について考察した．この結果を用いて，単純名詞句の範囲決定を行うことが今後の課題である．

なお，本研究は，一部文部省科学研究費補助金 (# 07558162) の援助により行った．

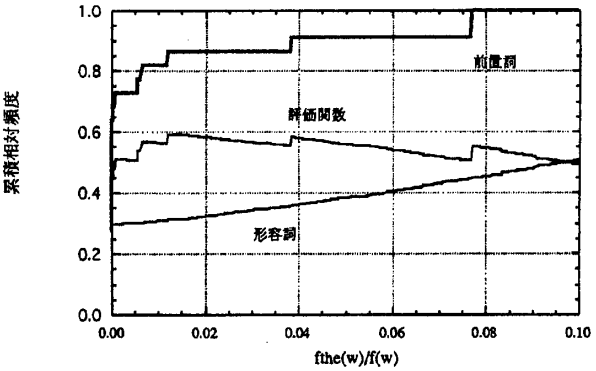

図 1 累積相対頻度と評価関数

表 3 低頻度語を除いたときの閾値 $t = 0.0119$ の結果

	$f_{\text{the}}(w)/f(w) > t$	$f_{\text{the}}(w)/f(w) \leq t$
$f(w) \geq 1$	161181	279717
$f(w) \geq 2$	135916	124817
$f(w) \geq 3$	119354	77869
$f(w) \geq 4$	108125	55308
$f(w) \geq 5$	99668	42225
$f(w) \geq 10$	76079	18687
$f(w) \geq 100$	26691	3643
$f(w) \geq 200$	19061	2717

表 4 低頻度語を除いたときの前方修飾語ではない語

	前置詞のみ	形容詞のみ
$f(w) \geq 1$	18	1433
$f(w) \geq 2$	18	944
$f(w) \geq 3$	18	703
$f(w) \geq 4$	18	564
$f(w) \geq 5$	18	499
$f(w) \geq 10$	18	275
$f(w) \geq 100$	18	72
$f(w) \geq 200$	17	56

参考文献

1) 竹田，松尾：英文科学技術文における単文の原子論理式への変換，情報処理学会第 49 回全国大会講演論文集 (1994).

2) 竹田，須田，楠本，松尾：英文科学技術抄録文における名詞の決定，情報処理学会論文誌 **36**(8), pp. 1828–1837 (1995).