英文科学技術文における前方修飾語の決定について

2C－8

丸木 健次† 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

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，原子論理式の述語記号に動詞を項に名詞句をそのまま単語列としてあてる方法が考えられる[1]．英文科学技術文の名詞句の範囲決定のために，まず一つの被修飾名詞[2]とその前方修飾語[3]からなる単純名詞句を決定する．以下の例文における下線を施した単語列が単純名詞句である．

> The values of the registration parameters are automatically calculated by maximizing an integer similarity measure selected for robustness.

名詞句は，単純名詞句が前置詞，接続詞などによって結合したものと考えることができ，単純名詞句は名詞句の基本単位であると考えられる．

単純名詞句は一つの被修飾名詞と前方修飾語からなる単語列である．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．そこで，本稿では前方修飾語の決定を問題とする．

2. 前方修飾語の決定

冠詞 the から始まる句はほとんど名詞句と考えられるので，the の次の語は前方修飾語であると考えられる．単語 w が現れる頻度を $f(w)$，'the' の後に現れる頻度を $f_{\text{the}}(w)$ とすると，$f_{\text{the}}(w)/f(w)$ は語の前方修飾語らしさの１つの指標と考えられる．この指標をこれ以降，前方修飾度と呼び，$m(w)$ で表す．以下の例文において [と] で囲まれた数は前方修飾度を示している．

The[0.000043] papers[0.036134] in[0.001669] this[0.000021] conference[0.163147] describe[0.000206] a[0.002217] wide[0.063583] range[0.177902] of[0.000011] applications[0.033038] of[0.000011] remote[0.108389] sensing[0.090566] methods[0.075211].

著者らは，前方修飾度が閾値 t に対して

$$m(w) > t$$

となる語を前方修飾語とする方法を提案している[3]．

閾値 t は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するようにして求めた．しかし，この方法で閾値を求めることには，前置詞の品詞のみをもつ語が少ないために問題がある．

そこで，本稿では前方修飾語と非前方修飾語のサンプルから閾値を求め，その評価を行った．

3. 閾値の求め方

閾値を求めるためには，前方修飾語と非前方修飾語のサンプルが大量に必要となる．通常このようなサンプルはコーパスから得られるのだが，目的とするような英文科学技術文のコーパスはなく，またその作成には多大な労力が必要となる．

そこで，制限を加えた疑似的な単純名詞句を抽出し，前方修飾語としてこの疑似単純名詞句の最終語以外の語を用い，非前方修飾語として疑似単純名詞句の冠詞の直前の語を用いることにした．例えば，以下の例で下線を施してある部分が単純名詞句であり，太字で記した語が前方修飾語のサンプル，斜体字で記した語が非前方修飾語のサンプルである．

... *of* the **registaration** parameters are ...

... *with* a **small air** track of ...

In all *cases* the **resonance condition** can ...

評価関数としては次を用いた．

$$\frac{m(w) > t \text{ となる } M\text{の数}}{M\text{の数}} \times \frac{m(w) \leq t \text{ となる} \neg M\text{の数}}{\neg M\text{の数}}$$

On Determination of Preceding Modifier in Scientific and Technical Documents

Kenji Maruki†, Masayuki Takeda† and Fumihiro Matsuo†
† Graduate School of Infomation Science and Electorical Engineering, Kyushu University, Hakozaki, Fukuoka 812-81, Japan

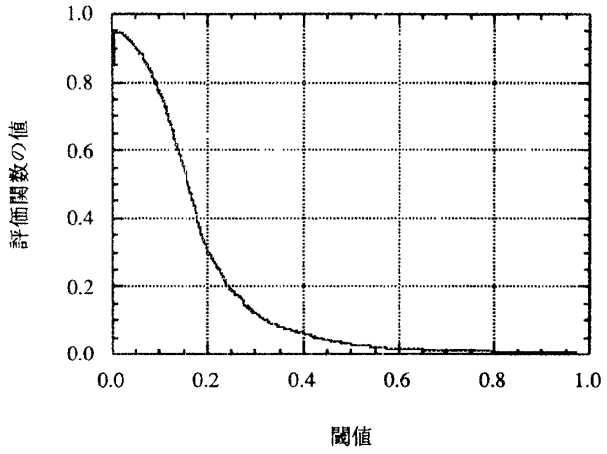

図 1 閾値と評価関数

表 1 閾値 $t = 0.007538$ の結果

	$m(w) > t$	$m(w) \leq t$	計
M	2,319,232	16,911	2,336,143
$\neg M$	39,927	921,387	961,314

この式中の M は前方修飾語のサンプルを，$\neg M$ は非前方修飾語のサンプルを表している．この評価関数を最大にする t を閾値とする．

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 文から前方修飾語 2,336,143，非前方修飾語 961,314 をサンプルとして抽出し，閾値を求めると $t = 0.007538$ であった（図 1）．このとき，$m(w) > t$ となる前方修飾語のサンプルは 2,319,232，$m(w) \leq t$ となる非前方修飾語のサンプルは 921,387 であった（表 1）．

この閾値 0.007538 を用いたときの結果として，失敗例である 2 つの場合をみてみる．

1 つ目は前方修飾語のサンプルのうち $m(w) \leq t$ となるものである（表 2）．この場合は，Monte Carlo の carlo のような固有名詞の一部，degrees, cm のような単位，a priori のような熟語の一部であるものが多くみられる．これらは直前に the のあることが少なく，$m(w)$ が小さくなっている．また，certain は冠詞をとる場合ほとんど a certain となるので $m(w)$ は小さい．a は不定冠詞ではなく記号や変数として用いられていたものである．

2 つ目は非前方修飾語のサンプルのうち $m(w) > t$ となるものである（表 3）．語 near, above, inside, outside は前置詞だけではなく形容詞としても用いられるため $m(w)$ が大きくなっている．また，paper, case, cases は以下の例文のように前置詞句の終わりで用いられていた．

表 2 $m(w) \leq t$ となる前方修飾語

単語	m(w)	数
certain	0.001122	1133
carlo	0.000133	1068
degrees	0.003831	836
dimensional	0.005060	722
a	0.002217	483
dependent	0.006834	444
priori	0.000819	405
nm	0.001535	394
cm	0.004277	380
mm	0.006993	337

表 3 $m(w) > t$ となる非前方修飾語

単語	$m(w)$	数
near	0.121475	3435
above	0.202170	1808
inside	0.052029	1024
paper	0.381080	547
case	0.449244	522
cases	0.081177	468
study	0.172981	362
given	0.023639	362
outside	0.116580	360
present	0.329226	348

In this paper a brief outline is ...

In all cases the resonance condition can ...

4. むすび

この論文では，前方修飾語の決定に用いる前方修飾度の閾値の求め方を示し，前方修飾度によって前方修飾語を決定することの問題点について考察した．

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．

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).

3) 丸木，柴田，日昔，竹田，松尾：英文科学技術文における単純名詞句の範囲決定，情報処理学会第 53 回全国大会講演論文集 (2), pp. 23–24 (1996).