英文科学技術文における基本名詞句の構造

7B－4

日昔 吉樹† 丸木 健次‡ 竹田 正幸‡ 松尾 文碩‡
†九州大学大学院工学研究科 ‡九州大学工学部

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]．ここでは，原子論理式の項となる名詞句の範囲決定に必要な基本名詞句に対し，機械的に擬似基本名詞句データを抽出し，このデータをもとに基本名詞句における単語間の結合について調査した．

2. 名詞句の範囲決定

ここでの名詞句の決定問題は，名詞句の構造を完全に決定するのではなく，名詞句を連続単語列としてその範囲を決定することである．

名詞は，形容詞によって修飾されるが，名詞によっても修飾される．そこで，例えばthe database systemのdatabaseを修飾名詞とよび，systemを被修飾名詞とよぶことにする．

基本名詞句の文法を図1に示す．基本名詞句とは，被修飾名詞とその前方修飾語のみからなる名詞句で，図1のNPである．以下の文において，下線を施した単語列が基本名詞句である．

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

名詞句の決定は次のような手順で行う．

(1) 被修飾名詞の決定．

(2) 一つの被修飾名詞とその前方修飾語からなる基本名詞句の決定．

$$NP \rightarrow N \mid Pr \mid M\ NP \mid Det\ NP$$
$$M \rightarrow Aj \mid N \mid Pr \mid Pa$$

N:名詞　M:修飾詞　Aj:形容詞
Det:決定詞　Pr:現在分詞　Pa:過去分詞
NP:基本名詞句

図1 基本名詞句の文法

(3) 基本名詞句をもとにした名詞句の範囲決定．

(1)の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．(2)，(3)において問題となるのは現在・過去分詞句による後方修飾，and, orの対応関係にある語の特定，現在・過去分詞による後方修飾の三つである．

3. 擬似基本名詞句データの抽出

多くの場合，基本名詞句の範囲は，被修飾名詞と動詞句，前置詞，冠詞などの単語によって決定できる．決定が困難となるのは主として前方修飾語がand, orで結合している場合や，基本名詞句中に現在・過去分詞を含む場合である．そこで，Dを冠詞，Nmを被修飾名詞とし，Eを前置詞，接続詞，関係詞の品詞をもたない語としたとき，正規表現DE^*Nmに合致する最大の単語列で，現在・過去分詞を含まないものを擬似基本名詞句と定義した．この定義に基づき，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文のうちand, orを含まない文を対象に，動詞句決定法[3]により決定した動詞句を除いた単語列から擬似基本名詞句を抽出した．その結果，冠詞を除いて2語以上の擬似基本名詞句は3,634,727抽出できた．このうち，冠詞を除いて2語の基本名詞句は約71%，3語の基本名詞句は約23%を占めている．

4. 基本名詞句の単語間結合

基本名詞句の決定の際に生じる曖昧さを解消するために，基本名詞句を構成する単語間の結合を調べた．

文の部分単語列がマルコフストリングであるというのは，それを単純マルコフ連鎖とみなしたときの生起

Structure of Simple Noun Phrase in Scientific and Technical Documents
Yoshiki Himukashi†, Kenji Maruki‡, Masayuki Takeda‡ and Fumihiro Matsuo‡
† Division of Engineering, Graduate School, Kyushu University 36, Hakozaki, Fukuoka 812-81, Japan
‡Faculty of Engineering, Kyushu University 36, Hakozaki,Fukuoka 812-81, Japan

表 1 基本名詞句の高頻度最終語
(相対頻度におけるかっこ内の数字は
非マルコフストリングのもの)

頻度	相対頻度	最終語
75089	0.0207 (0.0292)	system
66184	0.0182 (0.0202)	method
65735	0.0181 (0.0248)	model
43801	0.0121 (0.0110)	field
37400	0.0103 (0.0123)	structure
33363	0.0092 (0.0111)	range
28483	0.0078 (0.0085)	process
26890	0.0074 (0.0109)	problem
26785	0.0074 (0.0101)	function
26285	0.0072 (0.0065)	distribution
25073	0.0069 (0.0048)	dependence
24622	0.0068 (0.0049)	properties
23909	0.0066 (0.0064)	approach
22247	0.0061 (0.0049)	technique
22008	0.0061 (0.0058)	solution
21717	0.0060 (0.0047)	rate
21362	0.0059 (0.0059)	analysis
21164	0.0058 (0.0044)	equation
20773	0.0057 (0.0090)	effect
20630	0.0057 (0.0061)	region

確率p_mが，みなさないときの生起確率p_fより大きいものをいう[4]．

単語の生起確率と条件付き生起確率は，前述の10年分のINSPECテープ抄録文から算出した．ここで，単語とは大文字と小文字を区別しない英字と数字を字母とする文字列である．この抄録文の延べ単語数は，235,763,982，異なり語数は440,894であった．

まず，抽出した擬似基本名詞句がマルコフストリングであるかどうかを調べた．ただし，先頭の冠詞は無視した．その結果，約95.5%がマルコフストリングであることがわかった．このことを利用すれば基本名詞句の範囲決定の際に生じる曖昧さを減らすことができると考えられる．

疑似基本名詞句$w_1 \cdots w_n$において$p(w_n|w_{n-1}) < p(w_n)$となるものは，非マルコフストリング165,220のうち156,625あり，94.8%を占めていた．そこで，擬似基本名詞句の最終語に注目すると，例えば，systemは，3,634,727の擬似基本名詞句においては75,089回最終語として生起しており，非マルコフストリング165,220においては4,823回最終語として生起している．相対頻度はそれぞれ0.0207，0.0292となり，非マルコフストリングにおける相対頻度の方が高いことがわかる．表1に頻度の高い20の最終語についての擬似基本名詞句における相対頻度，()内に非マルコフストリングにおける相対頻度を示す．

例えば，語methodは，radio frequency methodやelectrical conductivity methodなどのように，基本名詞句radio frequencyやelectrical conductivityの後ろについて基本名詞句を形成する．このため，このmethodやmodel，systemなどを最終語にもつ句が新たに生成されると考えられる．語の生起確率は，過去の事例における相対頻度として算出しているので，上記のようにして新たに生成された句では最終語の直前の語に対する条件付き生起確率は低くなり，非マルコフストリングとなるものと考えられる．このことは擬似基本名詞句が非マルコフストリングになる最大の原因であると考えられる．

5. むすび

まず，抄録文から大量の擬似基本名詞句を抽出した．これをもとに，基本名詞句の単語同士の結びつきの強さを調べた．その結果，擬似基本名詞句の約95.5%がマルコフストリングであった．また，擬似基本名詞句が非マルコフストリングとなる原因の94.8%が名詞句の最終語の意味に依存することを明らかにした．このことを利用すれば基本名詞句の範囲決定の際に生じる曖昧さを減らすことができると考えられる．

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．

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) Nishimura, M., et al. : Determination of Verb Phrase in Scientific and Techinical Documents, Proc. Natural Language Processing Pacific Rim Symposium '95, pp. 95–100 (1995).

4) Takeda, M., Matsuo, F. : Markov String Grammar, Memoirs of the Faculty of Engineering, Kyushu University **55**(3), pp 279–284 (1995).