英文科学技術文における名詞句の決定について

1G－8

須田 淳一郎　竹田 正幸　松尾 文碩
九州大学工学部

1. まえがき

著者らは，英文科学技術抄録文理解システム作成のための第一段階として，専門語句の意味を無視した論理式への変換法を研究している[1]．本稿では，原子論理式の項となる名詞句の範囲の決定について述べる．

2. 名詞句の決定手順

ここでの名詞句の決定問題は，名詞句の統語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問題である．

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．この被修飾名詞の決定が，名詞句の決定において重要な役割を果たす．

単文の原子論理式への変換においては，まず動詞句決定法[2]により文の動詞句を決定し，次に名詞句を決定する．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成る基本名詞句の決定．

(3) 基本名詞句をもとにした名詞句の範囲決定．

3. 被修飾名詞の決定

被修飾名詞の決定については被修飾指数を用いた決定法[3]を開発している．その方法は，次のとおりである．単語wの被修飾指数(modificant index)$m(w)$を，表1のように定義する．いま，文が

$$w_1 w_2 \cdots w_n.$$

であるとすると，$m(w_i) - m(w_{i+1}) \geq 2$ならば，$w_i$を被修飾名詞とする．

この決定法を，人手により被修飾名詞を決定した17,128の名詞句に適用したところ，被修飾名詞の98%を決定でき，また，非被修飾名詞を被修飾名詞と誤認する割合は5%未満であった．

表1 被修飾指数

(a) 辞書にある語

前置詞(58語), 関係代名詞(5語), 接続詞(29語)	0
冠詞(the, a, an)	1
過去分詞形	1
現在分詞形	2
第1語義が副詞の語	0
第1語義が形容詞の語	1
第1語義が(代)名詞の語	4
上記以外で第2語義以降に(代)名詞をもつ語	2
上記以外の(助)動詞の原形，三単現，過去形	0
上記以外の語	4

(b) 辞書にない語

語尾がedの語	1
語尾がingの語	2
上記以外の語	4

4. 基本名詞句

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．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.

基本名詞句の決定における困難性は二つの場合に生じる．一つは，前方修飾語がand/orで結合している場合である．このうち，and/orの直前の語が形容詞の場合は，比較的容易である．すなわち，

> the combined electrical and thermal behaviour

では，electricalとthermalが対応関係にある．しかし，and/orの直前の語が名詞の場合には判定が困難である．例えば，

Identification of Noun Phrase in Scientific and Technical Documents
Junichiro Suda, Masayuki Takeda and Fumihiro Matsuo
Kyushu University 36, Hakozaki, Fukuoka 812, Japan

the velocity and temperature profiles

においては，velocity と temparature が対応関係にあり profiles を修飾しているが，この判定が難しい．

基本名詞句決定のもう一つの困難性は，前方修飾語に ing 形の語が出現する場合に生じる．例えば，

a single type of operating system

においては，operating は system の修飾語であり，operating system が基本名詞句である．一方，

a rapid and efficient method of processing telephone orders

では，基本名詞句 telephone orders は processing の目的語となっている．この問題に関しては，個々の ing 形の語について，その用法を調査中である．

5. 名詞句の決定

基本名詞句をもとにした名詞句の決定において問題となるのは，(1) 前置詞句の係り受け，(2) 現在・過去分詞句による後方修飾，(3) 連言問題の三つである．(2) は，現在・過去分詞句が後方から名詞句を修飾しているか否かの判定であり，現在，動詞ごとに用例を調査中である．(1),(3) については以下に述べる．

5.1 前置詞句の係り受け問題

前置詞句の係り受け問題では，次の三つを区別しなければならない．

(a) 前置詞句が名詞句を修飾している場合．

(b) 前置詞句が文の動詞に係る場合．

(c) 前置詞句が動詞とともに

change A into B

のような文パターンを構成し，A が B に変化することを意味しているような場合．

この三つの区別は，意味情報なしには困難にみえるが，実際には (b),(c) はあまり多くない．(b) の場合の前置詞句は，ほとんど 'in detail'，'at present' などのイディオムである．また，(c) の場合の文パターンをつくる動詞の種類はあまり多くないので[4]，文パターンについての規則の作成はさほど困難ではない．

5.2 連言問題

ここでの連言問題とは，and の対応関係にある被修飾名詞を特定する問題である．例えば，以下の文において，and の直後の被修飾名詞 location に対応する語が，and の前方の被修飾名詞 size, reduction, dependence のうちいずれであるかの判定は，意味情報なしには困難である．

The analysis also demonstrates the dependence of the achievable noise reduction on secondary source size and location with respect to the primary source.

そこで，各単語が被修飾名詞として生起する相対頻度 (これを被修飾度とよぶ) を，約 94 万文の疑似コーパスから算出し，これを用いた対応語の判定法を検討している．649 組の連言句について，対応する二つの語

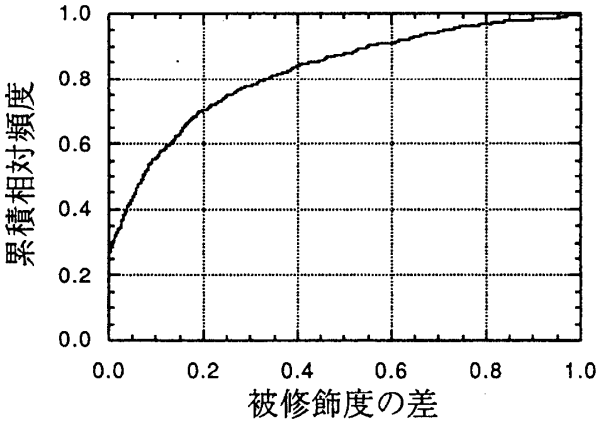

図 1 対応する語の被修飾度の差の分布

の被修飾度の差を図 1に示す．対応する語の被修飾度が近いものの割合が高いことがわかる．

6. むすび

本稿では，原子論理式の項となる名詞句の範囲決定について述べた．抄録文に現れる動詞の数は比較的少数であるため，用例調査や市販辞書に基づき，詳細な統語的・意味的規則を作成することができる．文の動詞句決定後にこれらの情報を用いることで，名詞句の決定において生じる曖昧さを解消できると考えている．

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，松尾：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 34(9), pp. 1931–1936 (1993).

3) 須田，竹田，松尾：英文科学技術文献抄録文における名詞句の決定，第 46 回電気関係学会九州支部連合大会講演論文集，p. 770 (1993).

4) 西村，竹田，松尾：英文科学技術文における前置詞を伴う動詞の研究，第 47 回電気関係学会九州支部連合大会講演論文集 (1994).