1N−2

英文科学技術用語の形態と統語

石鞍謙一郎　野中康吏　竹田正幸　松尾文碩
九州大学工学部

1. まえがき

著者らが現在研究を行っている英文科学技術抄録文理解システム[1]では，専門語句を意味ネットワークのノードとする．ノードとなる専門語句を完全に同定するには，専門語句を形成するすべての単語の構文・意味辞書をつくる必要がある．しかし，抄録文中に出現する単語の数は膨大であって，この作成には極めて大きな労力を必要とする．また，単語の異なり数は，延べ単語数の平方根に比例することもわかっている[2]．したがって，すべての単語について構文・意味辞書を作成することは不可能である．

抄録文中の専門語句は，機能語動詞の構文・意味情報により，かなりの確度で同定できると考えているが，もちろんこの方法は完全ではない．この方法を補うために，専門語句とわかっている句から抽出した形態的情報と統語情報を利用することが考えられる．

本稿では，専門語句として，2次文献情報であるINSPECテープの事前結合句243万(1989年1年分)を対象に，接辞および統語について調査した結果を述べる．また，機能語を含む専門語句について調査した結果についても述べる．

2. 単語の結合の強さ

high temperature superconductorでは，high temperatureがsuperconducterを修飾している．ここで，high temperatureは事前結合句中に128件出現するが，temperature superconductor，high superconductorは出現しない．このように，単語間の結合の強さを3単語の事前結合句中の各2語からなる事前結合句の出現頻度を比較することにより，3単語句中の2語の結合の強さを判定できると思われる．

ph_i^mをm単語からなるi番目の事前結合句，$f(ph_i^m)$をph_i^mの生起頻度，$N_m = \sum_i f(ph_i^m)$とし，3単語句$ph_i^3 = w_i^1 w_i^2 w_i^3$であるとする．ここで$w_i^1, w_i^2, w_i^3$は単語である．いま，

$$ph_j^2 = w_i^1 w_i^2, \quad ph_k^2 = w_i^2 w_i^3, \quad ph_l^2 = w_i^1 w_i^3$$

とし，

$$n_i = f(ph_j^2) + f(ph_k^2) + f(ph_l^2)$$

のとき，

$$p(ph_i^3[12]) = \frac{f(ph_j^2)}{n_i},$$

$$p(ph_i^3[13]) = \frac{f(ph_k^2)}{n_i},$$

$$p(ph_i^3[23]) = \frac{f(ph_l^2)}{n_i}$$

とし，それぞれ$ph_i^3[12], ph_i^3[13], ph_i^3[23]$の相対頻度(確率)という．$ph_i^3[12]$は$ph_i^3$の第1語と第2語からなる句と考える．$ph_i^3[13], ph_i^3[23]$についても同様に考える．また，$ph_i^3$の相対頻度$f(ph_i^3)/N_3$を$p(ph_i^3)$で表わす．

このとき，3単語句ph_i^3における2単語句の結合の強さd_{12}, d_{13}, d_{23}を次のように定義する．d_{12}, d_{13}, d_{23}は，それぞれ，第1語と第2語，第1語と第3語，第2語と第3語の結合の強さを表す指数とする．

$$d_{12} = \sum_i p(ph_i^3[12])p(ph_i^3),$$

$$d_{13} = \sum_i p(ph_i^3[13])p(ph_i^3),$$

$$d_{23} = \sum_i p(ph_i^3[23])p(ph_i^3)$$

調査の結果を，分野別に表1に示す．

表1　3単語句中の2単語句の結合の強さ

分野	物理学	電気工学 電子工学	制御工学 情報工学
3単語句数	612,294	298,895	287,215
d_{12}	0.280	0.270	0.261
d_{13}	0.139	0.146	0.153
d_{23}	0.432	0.379	0.357

次に，3単語句中の2単語句の出現が結合の強さと一致するかどうかについては，頻度の高い事前結合句における単語間の結合の強さを調査した．3単語句中の2単語

Morphology and Syntax of Scientific and Technical Terms
Kenichiro ISHIKURA, Yasushi NONAKA,
Masayuki TAKEDA and Fumihiro MATSUO
Faculty of Engineering, Kyushu University

句の出現は，*control system* synthesis などの $p(ph_i^3[12])$ が大きい事前結合句や，conputerised *picture processing* のような $p(ph_i^3[23])$ が大きい事前結合句は各2単語の結合の強さと良く一致する．しかし，*local area* network のように $p(ph_i^3[13])$ が大きい3単語句は，$ph_i^3[13]$ の第1語と第3語の間に単語を挿入してできた句を表わしていることになるが，実際の単語のかかり受けとは異なる場合が多い．

また，$n_i = 0$ であるような事前結合句は，この方法では単語間の結合の強さを判断できない．しかし，$n_i = 0$ である句を調査すると，*self adjusting* system, *time varying* system のように3語で一つの構造をなしているものはほとんどなく，いずれかの2語のかかり受けの構造をもつものがほとんどである．すなわち，3語の事前結合句は2語の事前結合句に修飾語を加えて構成されていると考えることができる．4語以上からなる事前結合句も同様である．

したがって，ほとんどすべての事前結合句は，2単語句に修飾を重ねることで作られると考えることができる．

3. 接尾辞

2節で述べたように，事前結合句は2語のかかり受け構造に基づく構文をとるが，単語単位で2語のかかり受け情報を辞書化することは，単語数が文献数の 1/2 乗で増加することから不可能である．そこで，接尾辞によるかかり受け構造を推定できるかどうかを調査した．接尾辞の個数は，一定しているのではないかと考えられるからである．最初に，2単語句について調査したが，いまのところはっきりとした結論が引き出せないでいる．そこで3単語句について調査を行なった．接尾辞として用いたのは ANCHOR 英和辞典の約1万語の常用語より抽出した 3,052 の最終音節である．

上に述べたように，3単語句 ph_i^m において $p(ph_i^3[23])$ が大きい事前結合句においては，第1語は $ph_i^3[23]$ を修飾していると考えられる．調査の結果，$p(ph_i^3[23])$ の大きい3単語句の第1語の語尾は，-y,-cal,-tic などの形容詞語尾が多く，$p(ph_i^3[12])$ の大きい3単語句の第3語の語尾は，-tion(s),-ment(s) などの名詞語尾が多い．よって，単語の語尾を専門語句同定の補助として用いることができる．

調査対象とした事前結合句に現われる単語 592 万語のうち，3,052 の接尾辞をもつものは全体の約 60%である．残りの単語は分野によりかなり偏りがみられ，この多くは分野の特徴を示す専門語であると考えられる．これらの単語については，接尾辞の情報が利用できないため分野別に辞書としてもつ必要がある．しかし，これらの単語数は，文献数が増加してもそれほど増えないと考えられる．

4. 専門語句中の機能語

対象とした 2,434,955 の事前結合句のうち機能語を含むものは，14.9%の 363,686 句である．今回は，機能語の前置詞について調査を行なった．特に，その前置詞が前置詞句の第1語であるかどうかを判断できるかについて調査した．

まず，to は他の機能語前置詞とは用法が異なり，1 to 1000 m , 1 kHz to 1 MHz のように範囲を示すものが 73%を占める．1986 01 26 to 1988 11 30 , 01 November 1978 to 11 December 1988 のように期間を示すものが 10%．残りも signal to noise ratio , mean time to failure などであり専門語句かどうかの判断は容易である．

その他の前置詞については，基本的に built in, on line のように成句として用いられるものと design for testability, analysis of covariance のように前置詞句として用いられるものがある．前置詞句として用いられる場合，専門語句の一部であるかどうかの判断は困難である．

5. むすび

専門語句の同定のための予備研究として，INSPEC テープの事前結合句を対象として，専門語句の接辞および統語について調査した．また，機能語前置詞を含む事前結合句についても調査を行なった．しかし，事前結合句と抄録文中の専門語句とは構文が異なっているので，この点についてはさらに調査を行う必要がある．

なお，本研究は一部文部省科学研究費補助金 (重点領域「知識科学」) によりおこなった．

参考文献

[1] 早田龍弘, 楠本典孝, 竹田正幸, 松尾文碩 : 英文科学技術抄録文における高頻度主題記述動詞の統語情報, 第 44 回情報処理学会全国大会, 1992.

[2] 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.