2Q－9

英文科学技術抄録文における高頻度動詞の格フレームに関する調査

河崎 裕司† 閻 仲天† 上福 潤‡ 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 ‡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]．この変換を行なうためには，文の動詞句の構造と名詞句の範囲を決定する必要がある．この変換の際に生じる曖昧さを解消するために格文法を用いることを考えている．しかし，英文科学技術抄録文において，どのような格フレームを用意すればよいのかわかっていない．そこで，EDR電子化辞書の専門用語辞書に含まれる格フレーム情報から出発して，これを実在の文書とつき合わせることにより，格フレームを精密化する方法が考えられる．本論文では，そのための予備的調査として報告する．

2. EDR電子化辞書の格フレーム

EDR電子化辞書の格フレーム情報は，情報処理分野に関する専門用語辞書に含まれる英語専門用語共起データにある．したがって，格フレームをもつ動詞は，情報処理分野に限られる．格フレームの一部を表1に示す．格フレームをもつ動詞の単語数は851，格フレーム生起数は3,236であった．

3. INSPECテープにおける動詞の調査

EDRの格フレームを精密化するには実在の文書とその格フレームをつき合わせ，問題点を顕在化させる必要がある．しかし，動詞の数は少なくないので，作業の効率化のために調査する動詞に優先順位をつけたい．そこで，文書集合としてINSPECテープを選び，動詞の頻度と分野による頻度の偏りを調べ，優先順位の参考にすることにした．

表1 EDR電子化辞書の格フレームの一部

integrate	D::object
integrate	S::agent
integrate	S::implement
integrate	into::goal
design	D::object
design	S::agent
design	for::purpose
design	with::manner

3.1 調査方法

1984年から1993年までの10年分のINSPECテープの抄録文を対象に動詞の生起頻度を調査した．INSPECテープとは，英国IEEが提供する英文二次文献データである．扱っている分野は，大きく物理学(INSPEC-A)，電気・電子工学(同-B)，制御・情報工学(同-C)，情報技術(同-D)に分けることができる．しかし，INSPEC-Dは文献数が少ないので，ここではINSPEC-Cに含めた．単文に関して動詞を高い精度で決定する方法[2]を開発しているので，これを用いて，動詞の生起頻度の調査を行った．調査対象を単文だけに絞るために，重文や複文の構成要素である接続詞，疑問詞，セミコロン，コロンなどを含まない文を疑似単文として抽出した．抽出された疑似単文は，約179万文であった．

3.2 情報処理分野の動詞

動詞決定法[2]を用いて決定した動詞の累積相対頻度を図1に示す．

EDR電子化辞書の格フレーム情報は，情報処理分野に関するものであるので，動詞の傾向が分野によっ

Preliminary Investigation for Case Frame of High Frequency Verbs in Scientific and Technical Documents
Yuji Kawasaki†, Yan Zhong Tian†, Jun Kamifuku‡, Masayuki Takeda† and Fumihiro Matsuo†
† Graduate School of Information Science and Electorical Engineering, Kyushu University, Hakozaki, Fukuoka 812-81, Japan
‡ Faculty of Engineering, Kyushu University, Hakozaki, Fukuoka 812-81, Japan

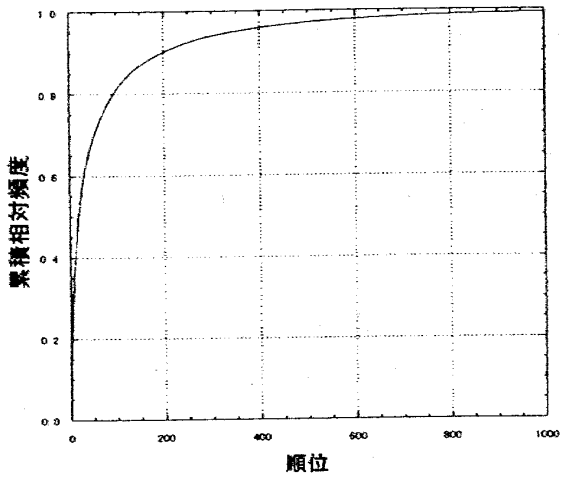

図 1　抄録文における動詞の累積相対頻度

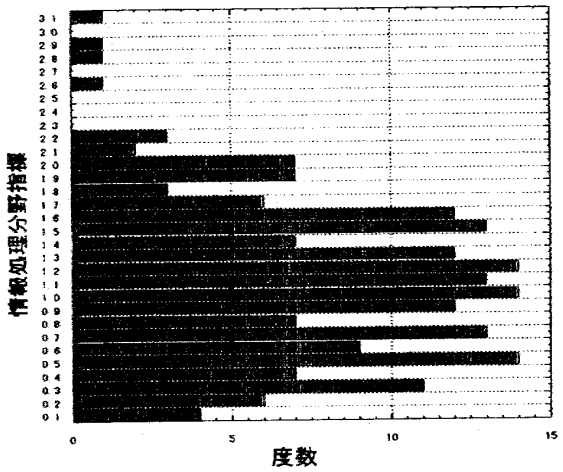

図 2　指標値の度数分布

て大きく異なるならば，格フレーム情報を適用する際に分野を考慮する必要がある．そこで分野による傾向を調査するために次の分野指標を考えた．

$$\frac{P^X(v)}{P^{A \cup B \cup C}(v)}$$

$P^X(v)$は，動詞vの文書集合Xにおける生起確率を表す．Xは，A，B，Cのいずれかであり，A，B，CはそれぞれINSPEC-A，-B，-Cの文書集合である．したがって、この指標は文書集合Xでの生起確率と文書集合$A \cup B \cup C$での生起確率の比を表わしている．この値が高ければXを特徴づける動詞と考えられる．

抽出した格フレーム情報は情報処理分野に関するものである．また，図1より抄録文に現れる動詞は，上位200語で全体の約90%を占めている．これらのことより，制御・情報工学・情報技術分野に関するINSPEC-Cに注目し，高頻度上位200語について上記の指標を求めた．指標値の度数分布を図 2に示す．指標が1.0前後付近には，take，know，playなどの日常一般によく使われる動詞が分布していた．また，指標の高い動詞の一部を表 2に示す．これらの動詞は，INSPEC-Cに特徴的な動詞と考えられる．

表 2　情報処理分野指標の高い動詞

単語	比	単語	比
run	3.0829	call	2.1948
implement	2.8487	address	2.1496
look	2.7779	integrate	2.0808
write	2.5817	help	2.0440
list	2.1971	define	1.9967

4.　格フレームをもつ動詞

INSPEC-Cに生起した動詞のうち，EDR電子化辞書に専門用語辞書の格フレーム情報が1つ以上ある動詞の単語数は660，格フレーム生起数は2,848であった．高頻度上位200語の動詞に限定すると，動詞の単語数は172，格フレーム生起数は1,019であった．

5.　むすび

本論文では，EDR電子化辞書にある格フレームを精密化するための予備的調査として，動詞の調査について述べた．

今後，実在の文書とつき合わせを行い，格フレームの精密化を行う．また，格フレームの精密化には，名詞自身からの格の決定の研究[3]も合せて考えている．

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

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) 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).

3) 河崎，丸木，上福，竹田，松尾：英文科学技術文における高頻度名詞の分類について，情報処理学会第55回全国大会講演論文集(2), pp. 208–209 (1997).