情 報 学 基 礎 25－8
（1992. 5. 12）

機能語による英文科学技術文献抄録文理解

竹田 正幸[1)]　早田 龍弘[2)]
石鞍 謙一郎[3)]　松尾 文碩[1)]

1) 九州大学工学部
2) 富士通
3) シャープ

原形約1,000語の機能語による英文科学技術文献抄録文理解の研究について述べる．高頻度主題記述動詞の用例を調査し，統語・意味情報を抽出したので報告する．また，動詞の優先度に基づく統語的曖昧さの解消法を提案する．

Understanding of Abstracts of Scientific and Technical Literature by Function Words

Masayuki TAKEDA[1)]　Tatsuhiro Sohda[2)]
Kenichiro ISHIKURA[3)]　Fumihiro MATSUO[1)]

1) Faculty of Engineering, Kyushu University
2) FUJITSU Corporation
3) SHARP Corporation

Understanding of abstracts of scientific and technical leterature by about 1,000 function words is discussed. Syntax and semantics about highly frequent verbs for describing subjects are reported. A method for resolving grammatical ambiguities based on verb priorities is also presented.

1 はじめに

著者らが研究している英文科学技術文献抄録文理解システムでは，専門的なことがらを表す語句(以後，専門語句という)をノードとし，ノード間の関係をラベル付きアークで与える意味ネットワークで抄録文を表現する．

自動索引の研究[2]によって，抄録文には現れやすいが専門語句には現れにくい語を統計的に取り出し，原形約1,000，原形+変化形約2,800語の機能語を得た．この機能語辞書の特徴的なことは，2,800語のうち65%の第一語義が動詞であることである．抄録文の延べ語数の半数がこの機能語辞書の語で占められており，この機能語によって区切られた単語列は，ほぼ専門語句とみなせる．これらのことは，機能語について詳細な統語・意味辞書を作れば，専門語句の抽出と，意味ネットワークのアークの自動ラベル付けが可能であることを示唆している．

機能語の動詞のうち抄録文に高い頻度で出現し，かつ論文の主題を記述すると思われる動詞を高頻度主題記述動詞と呼ぶことにする．五つの主題記述動詞 present, show, describe, study, discuss は，64%の抄録に出現する．これらの主題記述動詞の用例をINSPECテープ抄録文約11万を対象に調査し，統語・意味情報を抽出したので報告する．この結果は，情報システムにおける主題検索の高度化に利用できる．

自然言語においては，文は統語的な曖昧さをもっており，このことが自然言語理解の研究を阻む最大の障壁となっている．本稿では，文の骨格をなす動詞句を，文脈自由文法に基づく構文解析によらずに決定する手続きを提案する．この方法は，動詞に優先度を与え，複数の動詞候補が競合する場合に，この優先度に従って動詞句を決定するものである．この方法について，五つの主題記述動詞を含む849の単文を対象に評価を行い，96%の成功率を得た．この結果は，少なくとも抄録文については，統語的曖昧さを解消できることを示している．

2 高頻度主題記述動詞

抄録中の生起頻度順で上位40の機能語を表1に示した．主題記述に用いられると思われる動詞が，冠詞，前置詞などに混じって28～34位にかたまって現れている．これらの動詞を**高頻度主題記述動詞**と呼ぶ．INSPECテープ抄録文約11万を対象にこれらの動詞の用例調査を行い，統語・意味情報を抽出したので報告する．

五つの主題記述動詞のうち，discuss, describe は動詞以外の品詞はもたないが，study, show は名詞，present は名詞および形容詞の品詞をもつ．品詞別の割合は表2の通りである．表から show, shows, presents については，動詞としての用法がほとんどであり，したがって，抄録文においては，これらの変化形については動詞と考えてよい．その他のものについて，品詞を判別する手続きを与えなければならない．ここでは判別が最も複雑な present の原形について述べる．Present の形容詞用法には前位修飾・後位修飾のほかに叙述用法がある．前

表 1: 高頻度機能語 (上位 40)

1.	the	11.	on	21.	from	31.	describe
2.	of	12.	that	22.	or	32.	study
3.	be	13.	have	23.	result	33.	much
4.	and	14.	use	24.	model	34.	discuss
5.	a	15.	an	25.	can	35.	give
6.	in	16.	as	26.	author	36.	one
7.	to	17.	it	27.	these	37.	obtain
8.	for	18.	at	28.	present	38.	state
9.	with	19.	this	29.	show	39.	also
10.	by	20.	which	30.	used	40.	effect

表 2: 品詞別出現数

	動詞		名詞		形容詞	
present	714	(41.7%)	81	(4.7%)	917	(53.6%)
presents	800	(99.6%)	3	(0.4%)	—	
show	1779	(99.8%)	4	(0.2%)	—	
shows	990	(99.9%)	1	(0.1%)	—	
study	559	(27.9%)	1447	(72.1%)	—	
studies	80	(8.4%)	878	(91.6%)	—	

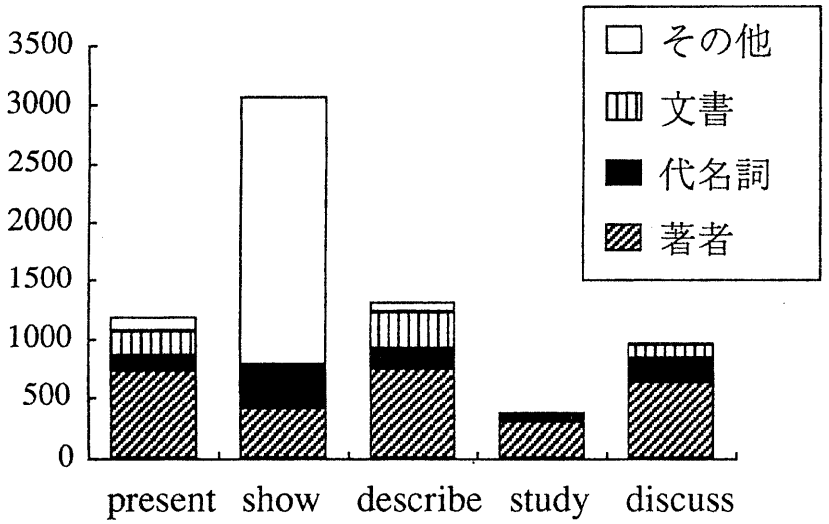

図 1: 主題記述動詞の主部

位用法は直前に冠詞や代名詞の所有格を伴うことが多く，叙述用法は直前あるいは副詞を挟んで be 動詞を伴うことが多いため，これらの判別は容易である．また，present の名詞用法は約 80%が at present であり残りもほとんどが the present であるため，動詞用法と間違う可能性は極めて低い．しかし，形容詞 present の後位用法ではこのような単純な判別ができず，被修飾語句を主部，present を動詞とする誤りの危険性がある．(present の動詞用法では後述するように主部が非常に限られているので，このことを用いれば判別は可能である．)

五つの主題記述動詞の動詞用法について，受動形は省き，動作の主体である主部を調査し，意味別に分類した．その例として，discuss の主部の分類を表 3に示す．

残り四つの主題記述動詞についても同様の調査を行い，分類した．その結果を図 1のグラフに示す．Show 以外の四つの動詞については，主部のほとんどは著者・代名詞・文書の三つである．そこで，これらの語句を伴う主題記述動詞の生起は，実際にこれらの語句を主部とする述語動詞であると考えてよい．したがって，統語的な曖昧さの解消に利用できる．また，著者・代名詞・文書の三つを主部とする主題記述動詞は，抄録の主題を記述していると考えられる．この結果は，情報検索システムにおける主題検索の高度化に寄与するものである．

3 統語的曖昧さの解消

文脈自由文法に基づく統語解析では曖昧さが生じ，文の統語構造を一意に決定することができない．このことが自然言語処理の研究を阻む最大の障壁となっている．この節では，文

表 3: discuss の主部

分類	句		割合
著者	the author(s)	the original authors	66.8%
	the present authors		
	Paine(1986)	Shingles	
	Spackman		
代名詞	they	it	20.1%
	he	she	
	we	one	
	few	these	
文書	the paper	this paper	13.1%
	the present paper		
	the article	this article	
	a resent article by Marin and Rosas(1984)		
	a later article		
	this presentation	the presentation	
	this report	the report	
	this review	the review	
	each review		
	this book		
	section 2	the following section	
	the third section	the last section	
	the fourth chapter	the first part	
	the program		
	this volume		

の骨格をなす動詞句を，優先度に基づいて決定する方法を提案し，その評価を行う．

3.1 優先度に基づく動詞句の決定

動詞の品詞をもつ語を V とするとき，the + V は明らかに動詞の候補から除外できる．このような単純な規則によって動詞候補を絞り込んだあと複数の候補が残された場合に，動詞の優先度に従って動詞句を決定する．優先度は次のように与えた．

優先度 1

(a) 助動詞を伴う動詞．

(b) 動詞以外の品詞をもたない機能語 (分詞を除く)．

(c) 特定の語句を伴う主題記述動詞．

優先度 2 機能語動詞 (現在分詞を除く)．

優先度 3 非機能語動詞 (現在分詞を除く)．

以下の例では，優先度 1 の動詞句を四角で囲み，優先度 2,3 のものをそれぞれ二重下線，一重下線で示している．

- The authors study the case of a symmetrical body immersed in a uniform flow at zero angle of attack.
- The efforts made for improving the Tsing Hua mobile educational reactor are described in this work.
- The article examines terminal use in a medium-to-large-scale network and presents an audit approach that can be tailored to the individual organization.

3.2 評価

あらかじめ単文とわかっている文については，文中の動詞句は一つなので，複数の動詞候補がある場合，優先度によって動詞を絞り込めばよい．しかし，実際は重文，複文の可能性があり，単純に動詞を一つに絞り込む訳にはいかない．現在のところ，あらかじめ重文，複文などの文構造を知る方法がないので，単文のみを対象に，優先度に基づく方法の評価を行う．

1989 年度配布の INSPEC テープから無作為に取り出した 329 抄録 (1376 文) に含まれる 849 の単文を対象に評価を行った．優先度ごとの動詞候補の平均出現数は，優先度 1 のものからそれぞれ 0.92, 0.50, 1.34 である．動詞候補数が 1 のものは 849 文のうち 204 文であった．動詞候補数が 2 以上ある 645 文について，優先度に基づく動詞句決定法を適用した．一

表 4: 成功内訳

優先度	種別	割合
1	助動詞を伴う動詞	61.4%
	動詞のみの品詞をもつ機能語 (分詞形を除く)	28.0%
	特定の語句を伴う主題記述動詞	9.8%
2	機能語動詞 (現在分詞形を除く)	0.8%
3	非機能語動詞 (現在分詞形を除く)	0.0%
合計		100.0%

表 5: 失敗内訳

原因	文の数	割合
優先度の低いものが動詞	3	8.6%
優先度 1 の動詞が複数	2	5.7%
優先度 2 の動詞が複数	20	57.1%
優先度 3 の動詞が複数	4	11.4%
助動詞との対応がとれない	1	2.9%
倒置文	2	5.7%
辞書にない動詞	3	8.6%
合計	35	

意に正しく動詞句を決定できれば成功，それ以外は失敗とする．その結果，645 文中 610 文が成功で，成功率は 94.1%であった．成功の内訳は表 4の通りである．助動詞を伴う動詞が 6 割強を占めており，そのほとんどが受け身であった．抄録文では，受け身表現が多いためにこのような結果となった．

次に，失敗となった 35 文について分類を行うと表 5のようになる．優先度 1 の動詞が複数個あったものは次に示す 2 文である．

- This [illustrates] a small section of a network comprising three service nodes each having a digital cross-[connect] system (DCS).
- As examples of service layer network components, two switches, an RCU, a 64 kbit/s cross [connect] and a flexible access system [are shown].

このように，辞書上では動詞以外の品詞をもたない connect が，名詞的に専門語句の一部と

して現れることがある．したがって優先度1の中にも段階を設けた方がよいと考えられる．倒置文の2文は，倒置文のパターンを考慮すれば容易に改善できる．辞書にない動詞は，いずれもformalizeが辞書にないために失敗となったものである．優先度2の動詞が複数個あって絞りきれない文が20と多いが，これはmodel, state, flowなど動詞として専門語句の一部として現れやすい語を考慮することによって，改善可能である．以上のように，失敗した文についても成功するように改善することは容易であり，実際には成功率をさらに高くできることがわかる．

4 おわりに

本稿では，原形約1,000語の機能語を用いた英文科学技術文献抄録文理解について論じた．五つの高頻度主題記述動詞についてINSPECテープの抄録文11万を対象に用例調査を行い，統語・意味情報を抽出した．この結果は情報システムにおける主題検索の高度化に寄与するものである．また，優先度に基づく動詞句の決定法を提案した．これにより，自然言語理解の研究の最大の障壁である統語的曖昧さが，少なくとも抄録文に関しては解消できる見通しがたった．

今後は，present, show, describe, study, discuss以外の主題記述動詞についても用例調査を行い，統語・意味情報を抽出する．また，動詞の優先度と接続詞などの情報により，重文・複文などの文構造を決定する手続きを確立するための研究を進めていく予定である．

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」)により行った．

参考文献

[1] 石鞍謙一郎，野中康吏，竹田正幸，松尾文碩: 英文科学技術用語の統語情報，第44回情報処理学会全国大会講演論文集(3)，pp. 93–94, 1992.

[2] 二村祥一，松尾文碩: 英文科学技術情報に対する不要語除去法による自動索引，情報処理学会論文誌，Vol. 28, No. 7(1987), pp. 737–747.

[3] 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講演論文集(3)，pp. 91–92, 1992.

１９９２年度情報学基礎研究会・シンポジウムの案内

科学における情報の円滑な流通と高度利用を促進する為，データ・知識に関する基本的問題とその整備・利用に関する研究交流を目的とした当研究会・シンポジウムを開催します。

今年度の題目は，最近注目されている研究題目の中から，次の様な題目を予定しています。論文・講演の申込み，及び参加の程，宜しくお願いします。

研究会の案内：

・論文申込み締切：　各研究会開催日の３カ月前
・原稿枚数（原則）：　８枚（Ｂ５）
・題目とスケジュール（予定）：
　－自己組織化　７月１４日（火），機械振興会館，Ｂ３－２
　－ゲノム　９月　８日（火），機械振興会館，Ｂ３－２
　－意味論　１１月１０日（火），機械振興会館，Ｂ３－１

シンポジウムの案内：

・講演申込み締切予定：　１９９２年９月中旬
・原稿枚数（原則）：　４～１０枚（Ａ４）
・題目とスケジュール（予定）：
　－オブジェクト指向と利用者インタフェース
　　１９９３年１月１３日（水），１４日（木），日本学術会議講堂（六本木）

問合せ先：

静岡県沼津市宮本１４０番地　〒４１０－０３
富士通㈱　情報システム事業本部　ＰＰ事業部　第一開発部
尹（ゆん）　博道
電話　直通（０５５９）２４－７２４０　代表（０５５９）２３－２２２２
ＦＡＸ　（０５５９）２４－６１９７