英文科学技術抄録文における動詞の決定

竹田正幸† 松尾文碩†

自然言語処理において，文の統語・意味構造の中心となる動詞の決定は重要である．本稿では，動詞候補間の優先度に基づく比較的簡単な方法によって英文科学技術抄録文の動詞が決定できることを示す．この方法を約2,400の単文に適用したところ，96.6%の成功率を得た．

A Method for Determining Verb of Sentences in Abstracts of Scientific and Technical Literature

Masayuki Takeda† and Fumihiro Matsuo†

The identification of the verb in a sentence is very important in natural language processing, as the verb plays a key role in recognizing the syntactic and semantic structure of the sentence. This paper presents a simple method for identifying the verbs in abstracts of scientific and technical literature. The method first selects some words as verb candidates from a sentence, arranges them in a priority order, and finally decides on the verb according to the order. This method applied to 2,400 simple sentences identified 96.6% of the verbs correctly.

1. まえがき

自然言語処理における統語解析では，1文に対して多数の統語的曖昧さが発生する．統語的曖昧さは，意味的制約によって解消することが一般的である[1]が，この制約の作成には多大の労力を要する．

文の統語・意味構造の決定において，中心的な役割を果たすのは動詞である．文の動詞を決定することができれば，動詞の統語・意味情報によって，文の統語的曖昧さを減少させることが可能になる．

本稿では，英文科学技術抄録文に対して，比較的簡単な方法によって動詞が決定できることを示す．その方法は，動詞の間に抄録文において動詞となるための優先順位を設定し，それに基づいて動詞を決定するものである．この優先度は，動詞を機能語と非機能語，常用語と非常用語に分けることによりあたえる．機能語とは，抄録文には現れやすいが専門用語（technical terms）には現れにくい原形約1,000語である．また，常用語とは，市販の学習英和辞典に重要語として記載されている語で，中学基本語約1,400語，高校標準語約5,600語を含む約12,500語である．優先度に基づく動詞決定手続きによって，約2,400の単文を対象に動詞の決定を行ったところ，96.6%の成功率を得た．

† 九州大学工学部電気工学科
Department of Electrical Engineering, Faculty of Engineering, Kyushu University

2. 機能語

機能語は，不要語除去法による自動インデキシングの研究[2]によって得られたものである．そこで得られた1,667語の不要語によって生成されるキーワードは，すぐれたインデキシング能力をもつ．この1,667語を変化形によって完備化すると，原形1,145語，原形＋変化形2,841語の辞書が得られる．この辞書の語を機能語という．

表1に，高頻度機能語上位100語を示した．機能語は，抄録文の延べ単語数において約半数を占め，**表2**に示したように原形＋変化形のうち65%の語の第一語義が動詞であるという特徴をもつ．

3. 動詞決定法

本稿の動詞決定法は，単文を対象にしたもので，まず文中の単語のうち動詞の品詞をもつ語を動詞候補として選び出し，次に各動詞候補に優先度を割り当て，優先度の最も高い候補を動詞とする方法である．そこで，この決定法は動詞候補選出手続きと動詞決定手続きから成る．

3.1 動詞候補選出手続き

動詞候補は，**図1**の統語規則によって選出する．この規則は，正規文法（右線形文法あるいは左線形文

表 1 高頻度機能語
Table 1 High frequent function words.

1	the	26	author	51	some	76	but
2	of	27	these	52	increase	77	consider
3	be	28	present	53	such	78	flow
4	and	29	show	54	given	79	may
5	a	30	used	55	their	80	calculate
6	in	31	describe	56	determine	81	only
7	to	32	study	57	found	82	made
8	for	33	much	58	into	83	small
9	with	34	discuss	59	large	84	point
10	by	35	give	60	both	85	investigate
11	on	36	one	61	test	86	compare
12	that	37	obtain	62	they	87	under
13	have	38	state	63	develop	88	single
14	use	39	also	64	provide	89	mean
15	an	40	effect	65	form	90	approximately
16	as	41	design	66	degree	91	during
17	it	42	find	67	observe	92	term
18	at	43	not	68	more	93	good
19	this	44	low	69	case	94	over
20	which	45	make	70	all	95	various
21	from	46	base	71	when	96	studied
22	or	47	well	72	change	97	will
23	result	48	new	73	other	98	each
24	model	49	its	74	include	99	possible
25	can	50	than	75	different	100	report

表 2 機能語の品詞
Table 2 Part-of-speech of function words.

第一語義による品詞	原形＋変化形	原形
動詞	1,851 (65.2%)	421 (36.8%)
名詞	334 (11.8%)	145 (12.7%)
形容詞	297 (10.5%)	228 (19.9%)
副詞	232 (8.2%)	224 (19.6%)
前置詞	37 (1.3%)	37 (3.2%)
代名詞	24 (0.8%)	24 (2.1%)
接続詞	20 (0.7%)	20 (1.7%)
補助動詞	14 (0.5%)	14 (1.2%)
冠詞	3 (0.1%)	3 (0.3%)
記号	29 (1.0%)	29 (2.5%)
計	2,841	1,145

法）に変換可能である．ただし，次の3種類の語は動詞候補とはしない．

- 直前に冠詞を伴う語．
- to 不定詞句中の語．
- 進行形でない現在分詞形．

図1の規則は，主として助動詞を伴った動詞句，特に“is also discussed”のように助動詞と動詞の間に副詞などが挿入された句を動詞候補とする規則になっている．また，“Also described are…”などの倒置形に対する規則もあたえている．上記禁止則のために必要なto不定詞句（*To-Inf*）や進行形（*PF*）の定義もここであたえている．

下記の二つの抄録文において，下線を施した語が動詞候補であり，斜体で示した語が動詞である．

文1：The calculated data *reveal* a strong dependence of the implantation and reflection feature on the incident angle, particularly at grazing incidence conditions.

文2：The optimization of this function with respect to the registration parameters *is performed* using an adaptive random search strategy.

この手続きによって，1989年度配布のINSPECテープのうち884文献の抄録から得た単文2,408文を対象に動詞候補を選出した．その結果，1文あたり平均約3個の動詞候補が選出されたが，2,408文のうち4文については，市販辞書にない語が動詞であったため，動詞候補とすることができなかった．それらは次の4動詞句である．

- was graft-polymerized
- is simply rederived
- are reinterpreted
- has been multitasked

市販辞書にない語が科学技術論文に出現する問題は，形態素解析によって解決することができる．形態素解析による品詞の決定は，本稿が扱っている問題とは別の問題であるため，ここではこの問題に深入りしない．

3.2 動詞決定手続き

動詞の決定は，動詞候補に優先度をあたえ，優先度が最も高いものを動詞とする．優先度のあたえ方は，次の考察に基づいている．

1) 動詞を伴う動詞候補が動詞となる確率が最も高い．
2) 次に高いのは，図1の *Be*（be 動詞）と *Have*（have 動詞）である．
3) その次に高いのは，機能語動詞である．機能語動詞のうちでは，市販辞書において動詞以外の品詞をもたない語は，他の機能語動詞より動詞となる確率が高い．

機能語動詞と非機能語動詞の出現度数を**表3**に示し

Vroot	: 動詞原形	*PE*	: 挿入句	*PF*	: 進行形
Va	: 過去形	*Be*	: be動詞	*PT*	: 完了形
Vtps	: 3人称単数現在形	*Have*	: have動詞	*AuxP*	: 助動詞句
Pr	: 現在分詞形	*Aux*	: 助動詞	*Inv-P*	: 倒置形
Pa	: 過去分詞形	*To-Inf*	: to不定詞句	*Vc*	: 動詞候補
Ly	: 語尾が-lyの語	*PV*	: 受動形	$	: 文頭

PE → already | also | always | even | ever | first | further | here | herein | however | in detail | in general | in turn | just | more | moreover | never | now | often | soon | still | then | thereby | therefore | thus | well | yet | *Ly* | ε.

Be → is | is not | isn't | are | are not | aren't | was | was not | wasn't | were | were not | weren't.

Have → have | have not | haven't | has | has not | hasn't | had | had not | hadn't.

Aux → can | can not | can't | cannot | could | could not | couldn't | do | do not | don't | does | does not | doesn't | did | did not | didn't | may | may not | mayn't | might | might not | mightn't | must | must not | mustn't | shall | shall not | shan't | should | should not | shouldn't | will | will not | won't | would | would not | wouldn't.

PV' → be *PE Pa*.

PF' → be *PE Pr* | be *PE* being *PE Pa*.

PT' → have *PE Pa* | have *PE* been *PE Pa* | have *PE* been *PE Pr*.

To-Inf → to *Vroot* | to *PV'* | to *PF'* | to *PT'*.

PV → *Be PE Pa*.

PF → *Be PE Pr* | *Be PE* being *PE Pa*.

PT → *Have PE Pa* | *Have PE* been *PE Pa* | *Have PE* been *PE Pr*.

Inv-P → $ *Pa Be* | $ also *Pa Be*.

AuxP → *PV* | *PF* | *PT* | *Aux Vroot* | *Aux PV'* | *Aux PF'* | *Aux PT'* | *Have To-Inf* | *Be* able *To-Inf* | *Aux* have *To-Inf* | *Aux* be able *To-Inf* | *Inv-P*.

Vc → *Vroot* | *Vtps* | *Va* | *AuxP*.

図 1 動詞候補 (Vc) の構文
Fig. 1 Syntax of verb candidate.

表 3 単文における動詞の区分
Table 3 Categories of verbs occurring in simple sentences.

動詞区分	助動詞 伴わない	助動詞 伴う	計
機能語	880	1,328	2,208 (91.6%)
¬機能語∧常用語	53	122	175 (7.3%)
¬機能語∧¬常用語	4	17	21 (0.9%)
市販辞書にない語	0	4	4 (0.2%)
計	937	1,471	2,408

た．この 2,408 文は，3.1 節で述べた単文である．動詞区分において，機能語とは，機能語のうち市販辞書で動詞の品詞をもつ 518 語である．また，常用語とは，市販辞書の常用語約 12,500 語のうち動詞の品詞をもつ 3,897 語である．両方に属する語は 478 語あり，したがって，機能語でない常用語動詞（¬機能語∧常用語）は 3,419 語となる．動詞区分で，¬機能語∧¬常用語とあるのは，機能語と常用語のいずれでもない動詞のことを指す．この表から，動詞の 91.6%が機能語であることがわかる．また，約 60%の動詞が助動詞を伴っていることもわかる．このことが上記考察の(1)と(3)の基礎になっている．

そこで，動詞候補に優先度を次のようにあたえる．

優先度 1：助動詞を伴う動詞候補．

優先度 2：*Be* (be 動詞) と *Have* (have 動詞)．

優先度 3：市販辞書上動詞以外の品詞をもたない機能語 (313 語)．ただし分詞形を除く．

優先度 4：動詞以外の品詞をもつ機能語 (203 語) およびすべての機能語動詞の分詞形．

優先度 5：機能語でない常用語動詞 (3,419 語)．

優先度 6：市販辞書上動詞の品詞をもつ語．

ただし，優先度 3 と 4 のあたえ方は，変化形によって異なる．例えば，show は動詞と名詞の品詞をもつので優先度は 4 であるが，その変化形 showed は動詞以外の品詞をもたないため優先度は 3 となる．また，analyse は動詞以外の品詞をもたないが，その変化形 analyses は，動詞と名詞の品詞をもつ．ここで，機能語動詞の分詞形を優先度 4 とするのは，分詞形は形容詞的に使われることが多いからである．ただし，現在分詞形は，動詞候補選出手続きにおいて候補から除外している．

動詞候補選出手続きに，上記の優先度を割り当てる手続きを付加したものを動詞優先度評価手続きという．この手続きを 3.1 節で示した抄録文に適用すると，以下のような出力を得る．

文 1：The calculated data reveal[3] a strong dependence of the implantation and reflection feature[4] on the incident angle[5], particularly at grazing incidence conditions[5].

文 2：The optimization of this function[5] with respect[4] to the registration parameters is performed[1] using an adaptive random search[4] strategy.

ここで [と] に囲まれた数字が優先度である．この

例では，優先度の最も高い動詞候補は，いずれの文も一つだけであり，それが文の動詞になっている．

4. 動詞決定法の評価

前章の動詞優先度評価手続きの出力において，最高優先度の動詞候補が一つで，それが文の動詞のとき，この動詞決定法が動詞決定に成功したという．したがって，不成功には，次の場合がある．

1) 最高優先度候補が動詞ではなかった．このとき，次の二つの場合がある．
   a) 低い優先度候補が動詞であった．
   b) 動詞句を誤認した．
2) 最高優先度候補が複数存在した．
3) 動詞候補でない語が動詞であった．

このうち，(1-a)が本質的な失敗である．(1-b)は動詞候補選出手続きの失敗であって，図1に示した統語規則により動詞候補として選出された語句が，正しい動詞候補ではなかった場合と，逆に動詞句であるのに句として動詞候補に選出されなかった場合の二つがある．(2)は，曖昧さであり，このような決定法では，曖昧さが生じるのはある程度避けえないので，優先度以外の手段によって，この中から決定しなければならない．(3)は動詞候補選出手続きの失敗であるが，この場合については3.1節で述べたように，ここでは(3)の場合を取り扱わない．

単文2,404文を対象に行った評価結果を**表4**に示す．成功率は，93.1%であった．失敗166件中，(2)の失敗が83.7%を占めている．優先度1の動詞候補が現れた1,467件のうち，失敗した1件は次の(1-b)の場合である．

There has been renewed[1] interest[4] in their application in differential gas[5] sensor arrays[5] and the association with cellular automata and neural networking methods.

表4 動詞の決定
Table 4 Determination of verb.

最高優先度	文	成功	失敗		
			1-a	1-b	2
1	1,467	1,466	0	1	0
2	297	297	0	0	0
3	336	331	2	0	3
4	268	133	21	0	114
5	35	10	3	0	22
6	1	1		0	0
計	2,404	2,238	26	1	139

すなわち，この失敗は“renewed interest”の過去分詞“renewed”を“has been renewed”と誤って動詞句と判断したことによる．

優先度2の動詞候補（have, be）が最高優先度である297文についてはすべて成功している．優先度3の動詞候補は，市販辞書上動詞以外の品詞をもたない語であるが，それが最も優先度が高くなっている336文のうち，5文が失敗している．(2)の失敗3文中2文と(1-a)の失敗2文中1文は，

This illustrates[3] a small section of a network comprising three service[5] nodes each having a digital cross[5]-connect[3] system (DCS).

のconnectのように，動詞が専門用語の一部として名詞的に使われていたので，このことが判別できれば成功になる．残りの2文は，

The article details[4] the advantages and disadvantages of T/Cs and pyrometers to help engineers[6] specify[3] the most reliable, economical, and flexible temperature measurement system of this type[5] possible.

のspecifyのように，to不定詞句中の原形不定詞を動詞候補としたため失敗した．この場合，“help engineers to specify”であることがわかれば，成功となるが，この検出は難しい．

動詞候補の最高優先度が4，5の場合，(2)の失敗が多く，このため成功率が低くなっている．特に，最高優先度が4の場合，268文中135文失敗しているが，そのうち114は(2)の失敗，すなわち，優先度4の機能語動詞候補が複数存在したことによる失敗である．したがって，機能語に関する統語情報を利用することにより，成功率を上げることができる．次章では，動詞決定法の改良とその結果について述べる．

5. 動詞決定法の改良

前章の動詞決定法の評価において，失敗の大多数は，機能語の最高優先度候補が複数競合するために動詞を一意に決定できないという失敗であった．この種の失敗は，優先度に基づく方法では完全には避けえないが，動詞決定法を次のように改良することにより，ある程度回避できる．

- 動詞候補選出手続きを改良し，動詞でない候補を可能な限り除去する．
- 優先度のあたえ方を変更し，動詞である候補の優先度が高くなるようにする．

競合した複数の候補のうち動詞でない候補は，次のいずれかである．

（A）動詞以外の品詞．

（B）過去分詞形の形容詞的用法．

（C）不定詞．

このうち，（B），（C）の判別は困難である．（B）には，前位修飾用法と後位修飾用法とがあるが，後位修飾用法は，過去形と過去分詞形が同形である場合には，取扱いが厄介である．すなわち，

被修飾語句＋過去分詞形

と

主語＋過去形

を見分けることが難しい．また，（C）には，原形不定詞の場合と，

to＋動詞原形＋…＋and＋動詞原形

のように直接 to を伴わない to 不定詞の場合があるが，いずれもその検出は難しい．

一方，（A）については，機能語に関する品詞判別手続きを適用することで，ある程度判別可能である．実際，機能語のうち，高頻度主題記述動詞 show, describe, present, discuss, study については，1989年度配布 INSPEC テープ抄録文のうち約 11 万文を対象とする調査で得られた統語規則により，95%以上の確度で品詞を判別することができる[3]．

品詞判別手続きは，二つの手続きから成る．一つは，単語の抄録文における品詞別生起頻度情報を利用するもので，語の生起のほとんどが一つの品詞で占められている場合に，その語の品詞を一つに限定する．ここでは，動詞か否かだけを判別できればよいので，次のようにした．

a）動詞以外の用例が稀な presents, show(s) を動詞と判定する．

b）動詞としての用例が稀な while, but, low, single, long, further, near を動詞以外の品詞と判定する．

この手続きは，品詞を決定すべき単語のみに依存するものである．もう一つの手続きは，前後の語句に依存するもので，機能語に関する統語情報を利用して品詞を決定する．ここでは，次のようにした．

c）in terms of, in detail, based on, as well as, with respect to, under study, at present, in place of, for instance, by means of における下線を施した語は，動詞ではない．

d）result, model, study に関する統語情報によりこれらの語の品詞を判別する．例えば，“experimental results” における results は名詞である．

e）著者を表す語句（the author(s), the present author(s) など）や主格の人称代名詞（they, he, she など）に続く動詞候補は，それらの語句を主部とする動詞である．

f）高頻度主題記述動詞のうち，用例調査で得た主部になりやすい語句[3]を伴った語は動詞である．

以上の品詞判別手続きによって動詞でないと判定された語は，動詞候補選出手続きにおいて候補から除去する．一方，動詞と判定された語には高い優先度をあたえる．ここでは，（a）には優先度3，（e），（f）には優先度1をあたえた．

改良した動詞決定法の評価を，前出の単文2,404文を対象に行った．その結果を**表5**に示す．今回の成功率は96.3%となった．表4と比較すると，成功した文の数は76増加している．優先度1の候補の増加数は193であるが，このうち新たに成功となった文の数は55であった．（2）の失敗の数は，139から66に減少している．機能語に関するより詳細な統語情報を利用すれば，（2）の失敗をさらに減らすことができると思われるが，これだけでは十分ではない．そこで，（2）の失敗の66文に対して，以下に示す方法により，候補を一つに絞ることを試みた．

- 直前に前置詞を伴う動詞候補を除去する．
- 過去形の動詞候補の直後が冠詞ならば，それを動詞とする．

これにより候補を一つに絞ることのできた文は8文あり，すべて成功であった．

この結果，最終的な成功率は，96.6%となった．

表5 改良した決定法に基づく動詞の決定
Table 5 Determination of verb based on improved method.

最高優先度	文	成功	失敗		
			1-a	1-b	2
1	1,660	1,659	0	1	0
2	295	295	0	0	0
3	269	265	1	0	3
4	145	85	18	0	42
5	34	10	3	0	21
6	1	1		0	0
計	2,404	2,315	22	1	66

6. むすび

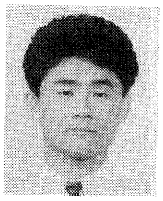
優先度に基づく比較的単純な手続きにより，非常に高い確度で単文の動詞が決定できることを示した．用例調査等によって得られる機能語についての詳細な統語・意味情報を利用すれば，この動詞決定手続きを更に改良することができる．この成果は，重文や複文における動詞決定にも利用できる．すなわち，動詞の優先度と接続詞，関係代名詞などの情報によって重文・複文などの文構造を決定することができると考えている．

なお，本研究は，一部文部省科学研究費補助金（#03245214，#04229216）の援助により行った．

参考文献

1) Allen, J.: *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Inc. (1987).
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, pp. 737-747 (1987).
3) 楠本典孝，竹田正幸，松尾文碩：英文科学技術文献抄録文における高頻度主題記述動詞，第45回電気関係学会九州支部連合大会講演論文集，p. 772 (1992).

(平成5年2月19日受付)
(平成5年6月17日採録)



竹田 正幸（正会員）

1964年生．1987年九州大学理学部数学科卒業．1989年九州大学大学院総合理工学研究科情報システム学専攻修士課程修了．同年4月から九州大学工学部電気工学科助手．パタン照合アルゴリズム，情報検索，自然言語理解などに興味をもつ．日本ソフトウェア科学会会員．

松尾 文碩（正会員）

1941年生．1966年九州大学工学研究科電子工学専攻修士課程修了．九州大学工学部電子工学科助手，同大型計算機センター講師，助教授を経て，現在同工学部電気工学科教授．工学博士．データベース，情報検索，自然言語処理，推論方式の研究に従事．