計算言語学 25－1
（1981. 2. 6）

専門用語の自動抽出

田中康仁 日本ユニバック㈱

はじめに

この研究は日本科学技術情報センターの第17回大会で最初の考えを発表した。また計量国語学会誌12巻8号に投稿している。このため一部内容が重複することがあるが，その後の研究内容を追加し，研究した成果を発表する。[31][32]

日本は原料の輸入により，それを高度に加工し，付加価値を付けて輸出することにより成り立っている。商品を輸出するにあたっても，ただ，ものを輸出するだけでなくこれに付属する情報を送らなくてはならない。例えばタービン一台を輸出するとすれば，約3000ページのマニュアルを翻訳しなければならない。[27]このように翻訳業務は社会的に重要な意味をもちはじめている。ここでは専門用語の自動抽出について考えてみる。

1. 専門用語の問題点

1－1 専門用語抽出の困難さ

専門用語を集めるにあたって困難なことは〝生物〟，〝医学〟等の専門分野で研究に携わっている人々は用語については関心あるが，自分達の研究に忙がしく専門用語を系統的に集め，整理するまでにはいたらない。一方，機械翻訳を実現したいと研究している人々は専門分野の知識を持っていない。このため専門用語を網羅的に収集することは困難である。専門用語の分析を行うには，日本語を分析し，その中から専門用語を抽出し対訳語を見つけ出すという方法がある。しかし，日本語は〝分かち書き〟といった問題があり，困難な面がある。ここでは英語の専門用語を分析し，その性質を利用して専門用語を抽出するという方法を考える。

1－2 長単位の用語について

これまでの用語の研究は辞書の編集とか情報検索のために行われてきた。このためには短単位を採用するほうが良い面が多かった。これは語の基本的意味が抽出できるとか，用語の数が多くならないとか，短単位の用語のほうが情報検索の検索に効果があるといった利点がある。しかし，機械翻訳の辞書を考えるにあたっては，長単位の用語辞書でなければならない。このことについてはこの研究会でも述べている。[33]

長単位の用語を集めることは収集の労力，件数の増大化が問題になる。また，どの程度の用語辞書を持てば，一つの専門分野の用語辞書として満足するかということも検討しなければならない。また，未知語，新しい専門用語についてどのようにして訳語を付けるかということも研究しなければならないテーマである。次に，これまでに行われている用語の研究について述べる。

2. 専門用語抽出の研究

用語抽出を考えるにあたって，まず考えることはKWICを作成し，用語を探す方法である。[1]

KWICを最初に提案した人はH·P·Luhnであり，1959年に発表している。KWICを作成する方法は単語をずらしながらレコードを作り，分類し，作成するものである。この方法は専門用語を見つけ出す一つの方法であるが，多量のデータを処理すると膨大なKWICが作成され，これを調べるだけでも大変な作業である。KWICを使用しない方法では九州大学の有川，武谷等が情報検索の研究から単語や用語の抽出を計算機を用いて網羅的に収集する研究を行っている。これは英文の中に現われる前置詞，冠詞，助動詞，代名詞，接続詞等をストップ・ワードとし，また別の特別の用語を準ストップ・ワードとし，これらストップ・ワード間の用語を専門用語とみて抽出している。しかし，これは経験的な方法であるため，大量に処理すると専門用語とは関係のない単語列を抽出する。このため，テーブルを追加するなどして防止しているが，これを系統的に防ぐ方策はなさそうである。だが単純なKWICによって用語を抽出する方法に比べると，はるかに優れた方法である。[2][3][4][15]

京都大学の長尾研究室ではタイトル文の機械翻訳という研究から専門用語抽出の興味ある方法を提供している。日本科学技術情報センターの抄録ファイルから，日本語タイトル，英文タイトルを同時に抽出し，その中から専門用語を抽出している。このようにすることによって〝電気〟，〝原子力〟といった分野の専門家でなくても，ある程度の訓練により対訳語を見つけることができる。この専門用語を見つける方法は単純なKWICでおこなわれている。[5][14][17]

その他の専門用語の収集，対訳語の作成は個々の企業や，協会，学会等で少しずつ行われている。これらの方法は既に発表されている各種資料や，辞典をもとにして作成されたものである。[8][7][10][24]

3. 専門用語の特徴分析

専門用語の特徴を分析し，収集，対訳語の作成のために，次のような分析を行った。

1) 調査対象データ　：　「学術用語辞典」電気工学編[7]　2) 調査対象件数　：　9,778件

3) 調査方法　：　専門用語の英語部分に品詞を付け，この品詞列がどのような構成になっているか調べた。この調査結果をまとめると，次のようになる。（次頁参照）

品詞列の種類ごとに集計すると品詞列頻度表のようになる。この表から専門用語は，名詞，形容詞の組合せで98.6％を占めている。また語数別分布を調べると，3語までがほとんどをしめ，4語以上のものも僅かながらあることがわかる。この分析作業でわかったこと，これら表から分析して得られることから専門用語抽出に役立つ事柄をまとめてみる。

1) 専門用語は名詞列であり，その名詞を修飾するために形容詞が使われる。このため品詞列から形容詞を無視して，できるかぎり長い名詞の品詞列を作り，その中から専門用語を探す。
2) 品詞列の最後の単語は名詞である。形容詞，その他の品詞が最後にはこない。
3) 5語以上の専門用語は発生する割合が非常に少ないので，5語以上は雑音として無視してもほぼかまわない。
4) 1語より構成する名詞
   1語よりなる名詞は非常に多く発生するため，基本的語（一般用語として使用頻度の高い語）

品詞列	件数	%
n	1,538	15.73
nn	4,349	44.48
an	2,438	24.93
nnn	537	5.49
ann	509	5.21
aan	80	0.82
nan	62	0.63
annn	47	0.48
nnnn	37	0.38
aann	15	0.15
nann	8	0.08
anan	8	0.08
nnan	6	0.06
aaan	2	0.02
naan	1	0.01
nnnnn	2	0.02
annnn	1	0.01
nannnn	1	0.01
その他	137	1.41
計	9,778件	100.00

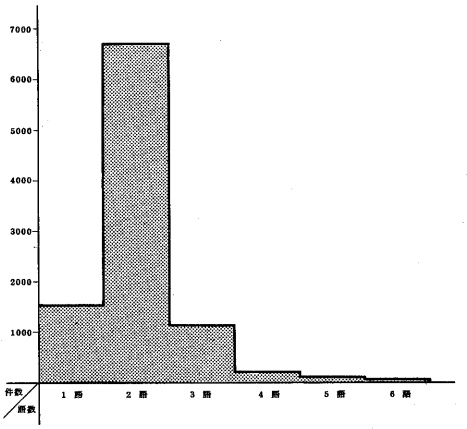
第1表　品詞列頻度表

（nは名詞，aは形容詞）

語数	件数	%
1語	1,538	15.96
2語	6,787	70.40
3語	1,188	12.32
4語	124	1.28
5語	3	0.03
6語	1	0.01
計	9,641	100.00 %

第2表　専門用語構成語数

"その他"の項目のデータを抜いたものである。



第1図　専門用語構成語数グラフ

を除去すれば専門用語だけが得られる。[11][12]

5) world, earth, sun などの前には定冠詞が付くが，このような特殊語の前の定冠詞は無視する。

6) 第1表の「その他」に区分された中を分析する。

名詞，名詞に形容詞が修飾したものをn'とし前置詞をprepで表示すると次のようになる。

① n' of n' 101件　② of以外の前置詞による形式 16件

③ 局所的andを含むもの 11件　④ 副詞⊕形容詞⊕名詞の品詞列 7件

⑤ その他 2件

7) 訳語の複数対応

専門用語はあいまいさが少なく一つの専門分野に対して一つの訳語が対応する。

しかし，ほんの僅かではあるが同じ電気工学の分野でも複数の訳語が対応する。

例1) 引込ケーブル ： entrance cable , leading−in cable , lead in cable , service cable , toll eutrance cable

例2) discarge voltage ： 放電電圧，制限電圧〔避電器〕

8) 用語の変化

用語は時代の移り変わりと共に少しずつ変化している。

例1) square wave ： 方形波(新)，矩形波(旧)

例2) アンテナ ： antena（米国式），aerial（英国式）

9) 略語，略記

非常に長い用語は，しばしば，省略形が使われる。

例1 OCR ： optical character reader 光学式文字読取装置

通常はOCRが使われている。

例2 current tap socket ： 分岐ソケット

current（電流）は省略されている。

日本語でも長い訳語が最初は用いられるが，しだいに短い略語が訳語となる。

10) 表記のゆれ

専門用語は表記方法が統一されているが実際に使用されている状況を調べてみると表記のゆれがみられる。

例 engineer ： エンジニア，エンジニヤ

これまで①～⑥で述べたことは，訳語の抽出に利用できる。 7)～10)で述べたことは非常に例外的なことではあるが専門用語辞書を維持，管理するうえで必要なことである。

4. 専門用語自動抽出

4－1 専門用語自動抽出プロセス

これまでに分析した内容にもとづき，一つの処理システムとして示す。

第2図の①～⑤とこのページの①～⑤の説明が対応する。

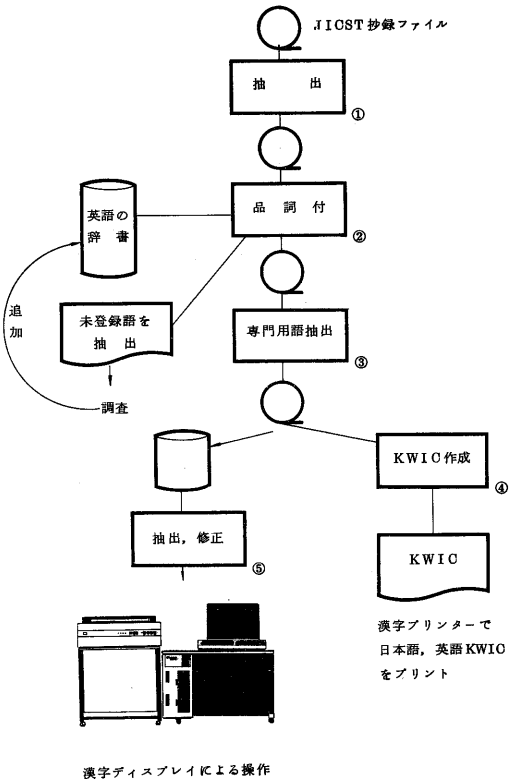
① JICST抄録テープより，日本文タイトルと英文タイトルを抽出する。または日本文と英文の対になったファイルを作成する。

② 抽出された英文に品詞を付ける。

品詞付けができない語をプリントし，調査の上，英語の辞書に追加する。

品詞付けの方法についてはもっと合理的な方法を研究中である。[28][9]

③ 専門用語になる可能性のある品詞列により用語を抽出する。



第２図　専門用語抽出プロセス

4）KWICを作成し，専門用語抽出の基礎資料を作る。

5）4）で出されたKWICより内容を調べ，漢字ディスプレイに表示し，同一訳語が対応すれば，この操作は省く。もし異なった訳語であるならばファイルに納める。

さらに具体的なデータにより専門用語を抽出する順序を示す。

①　JICSTファイルから英文，日本文タイトルの抽出

strategic planning approach to resource allocation。

資源割付への戦略的アプローチ

②　英単語に品詞付けをする。

strategic planning approach to resource
a　n　(n,v) prep　n
allocation.
n

③　用語の抽出

strategic planning approach
a　n　n
resource allocation
n　n

④　エディタと人間の介入により用語と対訳語の対応を付ける。（必要なら単語単位の形態素変形も行う。）

⑤　対訳語の確認

strategic planning approach ： 戦略的アプローチ
a　n　n
resource allocation ： 資源割付
n　n

4－2　専門用語抽出実験

専門用語抽出実験を4－1で述べた方法により行ってみた。

1）データ　：　JICST管理・システム技術編VOL 17.NO11　　2）データ件数　：　1,041タイトル

3）抽出された専門用語　：　3,002件　　4）1タイトル中の専門用語数　：　平均2.88個

5）1つの専門用語の平均単語数　：　1.98語

この方法にはまだまだ改良の余地があるが機械的に専門用語が抽出されることが確認された。

4－3　この方法の利点

1）漢字入力が省ける。　　2）専門的知識があまり必要でない。　　3）長単位の用語の抽出

4－4　専門用語と特殊語（数量詞）

専門用語は物質名，製品名を表現するものがかなりある。これら物質名や製品の量的多少を表現する形容詞が用いられる。これらは取り除いて処理するほうが望ましい。次にその単語を示す。

all, another, any, both, certain, each, either, enough, every, a few, less, more, much, neither, next, other, own, same, several, some, such

次に表示する形容詞はいくつかの専門用語で使われている。しかし，多くの場合は省いてもよいものである。

half, last, least, little, many, most, no, whole

例　least significant bit

これらは例外テーブルを用いて除外するようにすべきである。規則を作れば，必ず，例外が発生するので，例外については人間の判断が必要である。文章の中に現われるこれら単語の割合は 2.5 %程度である。

4－5　専門用語抽出（of や and を含むものについて）

専門用語の中には局所的 and や of 等の単語を含むものがある。これについて調べてみる。

①　of を含む専門用語

of を含む専門用語はいろいろあるので，これを統一的に抽出することはできにくい。of の含まれている部分を抽出し，手作業で調べるべきである。しかし，いくつかの特徴をもったものもある。

of を含むものは次のような語と結びついて大量に作られる。

例　table of ～，～表　：　time－lag of ～，～遅れ　：　angle of ～，～角
　　coefficient of ～，～係数　：　degree of ～，～度　：　density of ～，～密度
　　depth of ～，～深さ　：　duration of ～，～長　：　intensity of ～，～の強さ

これらについては特殊なテーブルを作成し，可能性リストをみつけ出すことができる。

専門用語は上にあげたような特別な語と〝of〟で結びつき専門用語を作るものと，分野特有の語と〝of〟で結びつき専門用語をなすものとがある。[7][8]

②　and を含む専門用語

and を含む専門用語を自動的に抽出するには意味カテゴリーの近いものを結びつけている and は局所的なものであると言われている。しかし，専門用語を抽出するために，抽出システムを複雑にすることはない。これらについては人間の判断にまかせるべきである。[29][30]

例1　scientific and technological information　：　科学技術情報
　　（scientific と technological は意味カテゴリーが近い。）

同一用語を省略している局所的 and もある。

例2　point and interval estimation　：　点推定と区間推定

例3　nondestructive testing and evaluation　：　非破壊検査及び非破壊評価

5. 専門用語の件数

一つの分野の専門用語はどの程度必要であろうかという疑問が起る。この問題に答えるために次の調査結果を示す。

1）調査データ　：　日本科学技術情報センター抄録テープ経営編[23]

2）抽出方法　：　文字種による漢字連続部分の機械的分かち書き

3）抽出データ　：　漢字部分のみ正確に50万件　　4）漢字列の種類：約8万件

漢字列の増加を5万件増すごとに漢字列の種類がどのように増加するかを調べた。この結果種類の増加はしだいに減少している。漢字列の中には次のような（上記物質，その他特性）正しくない結びつきもあるが専門用語の増加についてある程度の判断をすることができる。[33][18][19][20][21][22]

6. 専門用語と造語成分

専門用語を長単位で集めても限界がなく新語，新しい専門用語が生まれてくる。しかし我々はこれに実にうまい訳語を作っている。我々の知識の中に語に対する基礎的知識が有るからである。そこで専門用語を造語成分ごとに分解し，それに対応する訳語の造語成分を調べれば専門用語の造語成分辞書ができあがる。

例　屋内～　：　house ～，iuterior ～，indoor
substation〔電話〕
hyblid　：　ハイブリッド～，混成～，混合～
複合～

さらに英語の語根，接頭語，接尾語も上記のようにデータ・ベースにすれば新しい専門用語が出てきても，それにふさわしい訳語の組立ができる。例えばhyblid integrated circuitが新語と出てきたとし，hyblid，integrated circuit（集積回路）がわかっていれば〝混成集積回路〟〝複合集積回路〟，〝ハイブリッド集積回路〟〝混成集積回路〟という造語ができる。また既に使用されている用例を示すことによって複数の候補の中から一つを決めることができる。

調査漢字列件数	異なり漢字列種類	5万件ごとの増加種類
0 ～ 5万	14,202	14,202
～ 10万	24,280	10,078
～ 15万	32,929	8,649
～ 20万	40,518	7,589
～ 25万	47,785	7,267
～ 30万	54,988	7,203
～ 35万	62,045	7,057
～ 40万	69,084	7,039
～ 45万	75,527	6,443
～ 50万	81,594	6,067

第3表　漢字列調査件数と種類

新しい用語は新しい概念や事象の出現によって生まれる。これら概念や事象も改良され整理されて新しく変る。これにともなって用語も変わってくる。カメラ（Camera）の変化を示す。[26]

写真鏡（明4），照物　（明6），撮影箱（明21），暗箱（明41），写真機（昭3）撮影機（昭5）活動写真撮影機（昭7）

7. おわりに

専門用語は名詞句より構成されることは既知であり専門用語の語構成等も常識的なことではあると思われるがそれらを明確化し用語収集のシステムとして総合化することができたことは大きな意義がある。

謝　辞

この研究を進めるにあたって資料を提供して下さった九州大学の武谷助教授，京都大学の長尾教授，またJICST 抄録テープを提供して下さった日本科学技術情報センターの中井浩氏に深く感謝する。

参考文献

(1) 植村俊亮「電子計算機による自動索引の研究」(上) 電子技術総合研究所研究報告　第743号

(2) 武谷峻一，有川節夫「英文アブストラクトにおける単語の使われ方について」昭和54年度情報処理学会第20回全国大会

(3) 武谷峻一，宮野悟「REKWEST索引システムの改良について」九州大学理学部基礎情報学研究施設研究会

(4) Setuo Arikawa and Tosio Kitagawa "multistage Information Retrieval System Based upon Researcher Files." Kyushu university Research Institute of Fundamental Information Science.

(5) 長尾真「計算機による日本語文章の解析に関する研究」昭和53年度研究報告書

(6) 辻井潤一「ヨーロッパの言語処理の現状」　情報処理学会計算言語研究会資料21－1

(7) 文部省　学術用語集　電気工学編増訂版　電気学会

(8) マグロヒル「科学技術用語大辞典」　日刊工業新聞社

(9) Martin Lehnert　：　Reverse Dictionary of Present－Day English

(10) インタープレス　JISに基づく英和，和英"技術用語辞典"　インタープレス社

(11) 全英連　高校基本単語活用集　研究社

(12) Henry Kucera and W.Nelson Francis Computational anarysis of Present－Day－American English Brown University Press
John B.Carroll Peter Davies Barry Richman "Worcl Frequency Book" The american Heritage

(14) 長尾真他　：　日本語文献における重要語の自動抽出　　情報処理1976．2

(15) 武谷峻一，有川節夫，宮野悟　REKWEST 自動索引システムについて　昭和55年情報処理・学会第21回全国大会

(16) 水谷静夫　数学語表現の≪方言性≫　数学セミナー　79－9　日本評論社

(17) 長尾真，辻井潤一他　：　"国語辞書の記憶と日本語文の自動分割"　情報処理1978 VOL 19　NO 6

(18) 坂本義行　：　日本語テキスト分析のための語抽出実験(1)－特許公報の特定分野について－
第15回情報科学技術研究集会発表論文集

(19) 田中穂積　；「新編　日本語品詞列集成」　左順編（上下）　電総研

(20) 坂本義行　；「特許KWIC」　電総研

(21) 田島一彌，丹羽毓堂「日本語尾音索引」　笠間書院

(22) 風間力三　；「綴字順排列語構成による大言海分類語」　富山房

(23) 日本科学技術情報センター　：　JICST文献抄録テープ，経営編

(24) ドーランド　ドーランド医学大辞典　広川書店

(25) Allen B.Tucker，Jr Giuliano Gnugnoli，Long Vo Nguyen "Implemeutation consideraton for machine translation" 1978 ACM 0－89791－000－1　／78／0012／0884／

(26) 石綿敏雄，吉沢典男「外来語の語源」　角川書店

(27) 夕刊フジ"コンピューター「スペイン語←→英語翻訳シマス」"　昭和55年11月16日

(28) Jiirgen M.Janas "Automatic Recognition of the part－of－speech for English Texts" Information Processing & Managlment Vol 13 pp 205～213 Pergamon Press 1977

(29) 石綿敏雄　"言語の意味と言語情報処理"電子計算機による国語研究Ⅰ　国立国語研究所報告　秀英出版

(30) 石綿敏雄　"構文解析自動化の研究Ⅰ"　電子計算機による国語研究Ⅱ　国立国語研究所報告　秀英出版

(31) 田中康仁　"専門用語の解析と応用"　計量国語学12巻8号　計量国語学会

(32) 田中康仁　"専門用語の自動抽出"　第17回日本科学技術情報大会論文集

(33) 田中康仁　"漢字列長単位用語の抽出"　情報処理学会　CL研究会　'80年6月13日