情 報 学 基 礎 32－6
（1993．11．9）

文脈利用構文解析による意味情報の抽出とその応用

藤原譲　王暁晶　筑波大学
佐野彦麿　（株）トヨタ・テクノサービス

人間が技術文献などを読んで理解するプロセスは文章や句に含まれている意味情報を抽出するプロセスと言える。しかし，計算機などを用いて，構文解析だけで情報抽出を行なうと，その結果は漠然としていて明確で具体的な意味が分からない場合が多い。本報では，対象とする文章や句のまわりの文脈情報を利用することによって曖昧性を除くことを試みた。対象として分節索引と文献標題を選び，分野を限定したり分野特有の語を利用することにより，抽出結果に具体的な意味を与えることが可能なことを示した。抽出結果をファセット表にして表示することにより人間の理解を支援することが出来る。また，抽出語の意味カテゴリを利用することで構文構造のない非分節索引の情報抽出に利用できる。さらに，抽出結果の収集により知識獲得にも使える。用語集からの用語間の階層関係の自動抽出についても述べる。

Context-Dependent Syntactic Analysis for Extracting Semantic Information and Its Applications

Yuzuru Fujiwara, Wang Xiao Jing (University of Tsukuba)
Hikomaro Sano (Toyota Techno Service Inc.)

The human process of reading and understanding scientific or technical papers can be interpreted as that of extracting substantial semantic information involved in them. However, the meanings of the results extracted by syntactic analysis alone are often too vague to be readily understood by humans. This report describes a computer-aided syntactic analysis which utilizes the contexts around target sentences or phrases, in order to minimize the ambiguity and give the extracts concrete meanings. This approach was applied to articulated indexes and article titles. Facet tabulation of the extracts can aid humans in understanding the core ideas involved in documents. The semantic or facet categories assigned to extracted words are useful to semantic analysis of unarticulated indexes with no syntactic structures. Words extracted accordig to facet categories are also helpful to knowledge acquisition in the concerned subject. This report also describes automatic extraction of hierarchical relationships among terms from a new type index which comprizes complex technical terms.

1. まえがき

我々が技術文献などを読んで理解するプロセスは文章や句に含まれている意味情報を抽出するプロセスと言えよう。現在は発生する情報が多く，検索技術の進歩により検索は容易になったが検索で出力される情報も多い。これらの情報から機械的に意味情報を抽出し人間の理解過程を支援することが望まれよう。

文章などの解析・抽出には構文解析が用いられるが，一般的な構文解析だけでは抽出される情報の意味は広すぎて曖昧で直ちに人間の理解につながらない場合が多い[1,2]。そこで，対象とする文章や句の環境，すなわち，文脈を利用した構文解析をすることにより抽出情報に一義的で明確な意味を持たせることを試みた。本報では，更に，抽出結果の種々の利用可能性についても述べる。

2. 分節索引からの情報抽出

キーワード索引は科学技術文献の中心的内容をキーワードで示したものである。大量の文献をサーベイする場合には抄録などより短いキーワード索引の方が便利である。キーワード索引には，分節索引と非分節索引とがある。前者はキーワードを前置詞，接続詞などの機能語で結びつけキーワード相互の関係を明示的に示したものである[3]。後者は単にキーワードを羅列したものである。この分節索引から意味情報の抽出を試みる。

分節索引を採用しているデータベースには CA Search (Chemical Abstracts のデータベース版) や Analytical Abstracts などがある。CA Search のキーワード索引には，一般事項索引 (GSI)，化学物質索引 (CSI)，キーワード索引 (Keyword Index) がある[4,5]。前二者は分節索引で後者は非分節索引である。GSI からの情報抽出の例を述べる。

図1 GSIの例 (blood analysisの一部)

```
Blood analysis
  amikacin detn. in human, by HPLC, 21519j
  amiloride detn. in, by liq. chromatog., 84718u
  amino acid detn., in, of human by HPLC, aseptic
      meningitis in relation to, 251309v
  androstenedione detn. in, of women, by enzyme
      immunoassay, 127356u
  calcium detn. in, with ion-selective electrode,
      R 130827d
  clenbuterol detn. in, of cattle, by RIA, 47842j
```

図1はBlood analysisのGSIの例である。これから容易に分かる様に，detn.(determinationの略語) の前の語は血液中の物質名で分析の対象となる成分(analyte)を示している。"in"の次の語は血液試料を採る動物，または血液の部分を示しており，血液分析での分析対象試料(matrix)を表していると考えられる。"by", "with"の次の語は分析方法,分析機器(procedure)

を示している。最後の句で "by", "with"のないものは分析の目的，分析試薬，分析条件，影響因子など関連事項(note)を表している。さらに多くのGSIを調べた結果，そのパターンをBNF表記[6]で表したのが表１である。ここで非終端記号，A,M,m,Pは各々，分析成分，分析試料，分析方法で 血液分析における基本ファセットを示している。Nは関連事項，DTは文献の種類を示している。このBNF表記を利用すれば，各ファセット情報を抽出する抽出ルールをつくることができる。図１のGSIを対象に抽出をおこないファセット表に示したのが表２である。

表1　Blood analysisに関する GSI の BNF 表記

```
      <IM>::=<A><detn.>in<m>,<,><by><P>,<N>,<DT><DN>

      <IM>::= words for index modification
       <A>::= words for analyte substance
    <detn.>::= detn. | anal. | extn. | sepn. | assay|
               identification | monitoring | screening
   <M>,<m>::= words for blood part or animai | nil
      <in>::= in | of | nil
       <,>::= , | nil
      <by>::= by |with | using | nil
       <P>::= words for analytical procedure | nil
       <N>::= words for related matter | nil
      <DT>::= R | B | P | nil
      <DN>::= document number + check letter
-------------------------------------------------------------
      detn. = determination     anal. = analysis
      extn. = extraction        sepn. = separation
      R, B, P = review, book, patent
```

表2　Blood analysisのGSIのファセット表

Analyte	Matrix	Procedure	Note	DT	No.
amikacin	human	HPLC			21519j
amiloride		liq. chromatog.			84718u
amino acid	human	HPLC	aseptic meningitis		251309v
androstene-dione	women	enzyme immunoassay			127356u
calcium		ion-selective electrodes		review	130827d
clenbuterol	cattle	RIA			47842j

表３は他の GSI とその関連ファセットの例を示す。これらからも上記と同様にファセット情報の抽出ができる[7,8]。Analytical Abstracts の分節索引にも適用できる[9]。

表 3　GSI索引とその関連ファセットの例

GSI headings	Facets
Anion exchangers	Anion, Exchanger
Chlorination	Chlorinating agent, Chlorinated compound
Coating process	Substrate, Coating substance
Contact angle	Overlying liquid, Underlying solid
Corrosion inhibitors	Inhibitor, Protected material, Environment
Diffusion	Diffusing species, Matrix
Epitaxy	Substrate, Growth layer
Fireproofing agents	Fireproofing agent, Protected material
Heat of adsorption	Adsorbate, Adsorbent
Permeability	Permeating species, Matrix
Plasticizers	Plasticizer, Plasticized material
Toxicity	Toxic substance, Affected organism or organ

3．　文献標題からの情報抽出

分析化学に関する文献の標題には "determination","analysis"など分析の意味を表す語（ここでは分析関連語という）が含まれていることが多い。血液分析に関する文献の標題にも分析関連語が含まれてる。その標題の例をBNF表記とともに示したのが表４である。実際にはもっと多くのパターンが考えられる[10]。表４で非終端記号の A,M,P は各々，分析成分，分析試料，分析方法で，これらは分析の基本ファセットである。これらの文献標題についても各ファセット情報を抽出することができる[11]。

表４　血液分析に関する文献表題の例とそのBNF表記

Determination of theophylline in serum by HPLC.
<detn><A><in><M><by><P>

Direct determination of lead in whole human blood by total reflection x-rayfluorescence spectrometry.
<P><detn><A><in><M><by><P'>

Electrochemical determination of azidothymidine in human whole blood.
<P><detn><A><in><M>

Immunoenzymatic method for the determination of antimicrosomal auto-antibodies in blood serum.
<P><method><for><detn><A><in><M>

表4　続き

```
<detn>::= determination of | determining | analysis of | analysing |
          ····
<method>::= method(s) | procedure(s) | detector(s) | ····
<in>::= in | by
<by>::= by | using | by using | ····
<for>::= for | of
<A>::= words for analyte
<M>::= words for blood part or animal
<P>,<P'>::= words for analytical procedure
```

4.　用語集からの階層関係の抽出

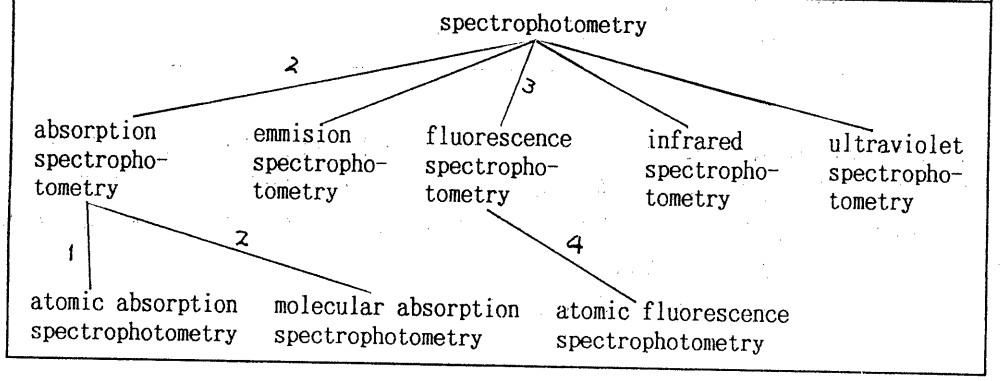
複合語はその構成要素の語より specific な概念を表す。日本語でも英語でも前方の修飾語は後方の被修飾語の概念を狭く限定する。こうした規則に基ずいて用語を修飾語数の昇順に並べたインデックス（SS-KWIC, Semantically Structured Key Word element Index in Terminological Context）を作成した。このインデックスからは自動的に階層関係が得られ，かつ用語間の定量的な類似度も得られる（表5参照）[14]。

表5　用語からの階層関係の抽出

SS-KWIC

spectrophotometry
absorption spectrophotometry
emmision spectrophotometry
fluorescence spectrophotometry
infrared spectrophotometry
ultraviolet spectrophotometry
atomic absorption spectrophotometry
molecular absorption spectrophotometry
atomic fluorescence spectrophotometry

階層関係



5. 抽出結果の利用

5.1 ファセット表示

表2の様に抽出結果をファセット表にして表示することで見やすく理解しやすくなる。人間が，ある物事を理解することは，それに関連したファセットに具体的な値を入れることであると言われているので[12, 13]，ファセット表示は人間の理解過程を支援することになる。

5.2 非分節索引解析への応用

2および3で述べた方法を用いると抽出語にファセット・カテゴリを自動的に付与できる。表6は血液分析に関する GSI からの抽出語の例である。

表6 Matrix, procedure, note カテゴリに
属する CA, GSI からの抽出語の例

Matrix	detection	radioassay
fish	detn	radioenzymic
human	device	RIA
newborn	extn	sensor
plasma	electrochem	spectrometry
platelet	ELISA	tomog
serum	enzyme	
vessel	fluorometry	Note
	gas(chromatog)	
Procedure	HPLC	book
	HPTLC	comparison
absorption	immunoassay	control
agglutination	IR	diabetes
analysis	liq(chromatog)	forensic
assay	mass	quality(control)
centrifugal	microcolorimeter	review
chromatog	photon	
	phosphorimetry	

各カテゴリに属する語数は matrix, procedure, note では比較的少なく analyte は多い。これを用いると構文構造のない非分節索引の解析ができる。図2は血液分析についての CA のキーワード索引である。これに表6を用いて各語をファセット毎に分けファセット表にしたのが表7である。表6に含まれない語は analyte とみなした。

また，ファセット・カテゴリは不規則構造の索引や文献標題の解析にも使える。

図2　CA,キーワード索引の例

Blood
cyclosporine detn immunoassay HPLC
RIA 21513c
furosemide detn HPLC forensic 23821a
plasma analysis gas chromatog review
130787r
sarafloxacin detn fish HPLC 126286j
serum pyrazinamide HPTLC 126293j
vessel calcium detn review 45287h

表7　CA,キーワード索引のファセット・テーブル表示

Analyte	Matrix	Procedure	Note	No.
cyclosporine		immunoassay, HPLC		21513c
furosemide		HPLC	forensic	23821a
	plasma	gas chromatog	review	130787r
sarafloxacin	fish	HPLC		126286j
pyrazinamide	serum	HPTLC		126293j
calcium	vessel		review	45287h

5.3　知識獲得

例えば，procedure として抽出された語を収集することで現在どんな分析方法が使われているかが分かる。この様に各カテゴリに属する語を収集することで知識獲得ができ，それをシソーラスの形で表すことができる。

5.4　キーワードへのロール付与

抽出語にロール（role）を付与することによりロール検索が可能になる。

5.5　シソーラス作成

4で述べたように用語間の上位，下位，同位関係が得られ，情報検索用のツールであるシソーラスが作成できる。

6. あとがき

文脈を利用したり分野を限定することにより構文解析が簡単になり，具体的で明確なファセットを設定できる。ファセット毎に自動抽出した語を利用することにより，人間の理解過程の支援，不規則構造の索引，標題，構文構造のない語列に対する抽出，知識獲得，ロール付与，シソーラス作成などが可能になる。こうした抽出された意味情報，および，その集積ファイルはさらに他のファイルと統合されて情報の高度処理システムの構築に役立とう。

7. 参考文献

1) N. Sager, L. Hirschman, "Information Structures in the Language of Science: Theory and Implementation" PB Report PB-289949 187p (1978)
2) ハリー・テナント著，森健一，河田勉，天野真家訳 "自然言語処理入門" 産業図書 (1984)
3) 川村敬一 "Articulated Subject Indexの構造的特性と機械化プログラム" 情報管理 24(5) 447-456 (1981)
4) 小川雅彌代表監修 "化学文献の調べ方" 化学同人 (1983)
5) Hedda Schulz著 吉田政幸訳 "CASオンライン入門" 地人書館 (1989)
6) 木村泉，米澤明憲，"岩波講座 情報科学－12 算法表現論" 岩波書店 (1982)
7) H. Sano, "Facet Tabulation of Index Terms" Information Processing & Management 26(9) 543-548 (1990)
8) 佐野彦麿 "分節索引のテーブル表示" 第29回情報科学技術研究集会発表論文集 P.353-359 (1992)
9) H. Sano, "Concise Display of Index Entries" Journal of Documentation 47(1) 23-35 (1991)
10) H. Sano, Y. Fujiwara, "Syntactic and Semantic Analysis of Article Titles in Analytical Chemistry" Journal of Information Science 19(2) 119-124 (1993)
11) H. Sano, "Extraction of Facet Terms from Article Titles and Their Display in Tabuular Form" Journal of Information Science 17(1) 43-48 (1991)
12) M. Minsky, "A Framework for Representing Knowledge" The Psychology of Computer Vision, edited by P. Winston, McGraw-Hill (1975)
白井良明，杉浦厚吉訳 "コンピュータービジョンの心理" 産業図書 (1979)
13) 浮田輝彦，木下聡，"自然言語理解のための知識表現と推論 － 知識表現" 情報処理 30(10) 1224-1231 (1989)
14) Y. Fujiwara, J. Lai, T. Makino, "Management and Advanced Utilization of Semantically Organized Terminology and Knowledge" Proceedings of the 3rd International Congress on Terminology and Knowledge Engineering p.141-151 (1993)