情 報 学 基 礎 38－3
（1995．7．28）

科学技術用語オントロジーの自動作成

松尾文碩　柴田誠　竹田正幸
matsuo@ee.kyushu-u.ac.jp
九州大学工学部
〒812-81 福岡市東区箱崎6-10-1

オントロジーのような語彙体系は，自然言語処理においても必要となる場合が多い．一般に，オントロジーの作成は多大の労力を必要とする困難な作業である．また，科学技術分野では，語彙はたえず増加しているため，困難性が更に大きい．本稿では，英文科学技術文を対象にオントロジーの自動作成について，基礎的考察を述べ，それに基づきINSPECテープの5年分を対象に行った実験結果を報告する．

AN APPROARCH TO AUTOMATIC CONSTRUCTION OF ONTOLOGY ABOUT SCIENCE AND TECHNOLOGY

Fumihiro Matsuo, Makoto Shibata, and Masayuki Takeda

Faculty of Engineering, Kyushu Univershity
Hakozaki, Fukuoka, 812-81 Japan

Ontology as a vocabulary system is important for natural language processing. Constructing the ontology generally requires a great effort. Especially, the ontology about sience and technology is dificult to make, because the vocabulary increases continuously. This paper presents an approarch to automatic construction of the ontology about science and technology.

1．まえがき

オントロジー (ontology) は，もともと哲学用語であり，イデア的存在 (ta onta) に関する理論を意味する．しかし，人工知能 (AI)，特に知識工学 (knowledge engineering) では，それは AI システムにおける語彙の体系と考えられている．本来のオントロジーと区別するために，AI オントロジーということもある．

オントロジーのような語彙体系は，自然言語処理においても必要となる場合が多い．一般に，オントロジーの作成は多大の労力を必要とする困難な作業である．また，科学技術分野では，語彙はたえず増加しているため，困難性が更に大きい．本稿では，英文科学技術文を対象にオントロジーの自動作成について，基礎的考察を述べ，それに基づき INSPEC テープの 5 年分を対象に行った実験結果を報告する．

2．順序集合としてのオントロジー

Dahlgren と McDowell は，生物学の分類 (taxonomy) が木構造であるのに対し，オントロジーはサイクルのない有向グラフとした[1]．彼らの定義は，オントロジーが生物学的分類とは異なり，複数の上位概念をもつとした Rosch 達[2]の説に基づいている．それならば，オントロジーの数学的定義としては，順序集合の方が適切かつ自然であろう．

オントロジーを語彙の順序集合と考え，語彙の類似性を位相数学の開集合によって規定した場合の数学的性質を文献 3 であたえた．この節では，この形式化を紹介し，問題点を述べる．

2．1　フィルター型内包演算

オントロジーとは，有限順序集合 $(X,\preceq)$ を意味する．すなわち，Xは，有限集合で，任意の $x,y \in X$に対して$\preceq$ は，$x \preceq x$; $x \preceq y \wedge y \preceq z \Longrightarrow x \preceq z$; $x \preceq y \wedge y \preceq x \Longrightarrow x = y$を満たす．

オントロジーに類似性あるいは近さという概念を導入するために，オントロジーに位相を定義する．$(X,^\circ)$ が位相空間であるというのは，Xが集合であり，$^\circ$が $A,B \subseteq X$に関し，次の (i)~(iv) を満たしているときである．

(i) $(A \cap B)^\circ = A^\circ \cap B^\circ$;　(ii) $A^\circ \subseteq A$;　(iii) $(A^\circ)^\circ = A^\circ$;　(iv) $X^\circ = X$.

Xの部分集合についての演算$^\circ$は，内包演算 (interior operation) と呼ばれる．A°を A の開核 (open kernel) と呼ぶ．$A^\circ = A$ のとき，A は開集合という．(iv) と (ii) から，Xと$\emptyset$ は開集合である．任意の $x \in X$について，x の近傍とは，$x \in A$ であるような開集合 A のことである．x の近傍の要素は，x に近いと考えられる．次の性質は容易に導かれる．

(vi) 任意個数の開集合の和集合は開集合である．

オントロジー $(X,\preceq)$ についての内包演算を定義するに当たっては，近傍と類似性の関連を考慮に入れなければならない．例えば，あるオントロジーにおいて，**犬** $\preceq$ **哺乳類**，**猫** $\preceq$ **哺乳類**，**馬** $\preceq$ **哺乳類**，... であるとする．この場合，集合 $\{x \mid x \preceq$ **哺乳類** $\}$ は，**犬**，**猫**，**馬**，... の近傍と考えることができ，また**犬**は，**猫**と**馬**に似ていると考えることができる．そのような集合を順序集合のフィルターと呼ぶ．もっと正確にいえば，オントロジー $(X,\preceq)$ において，$x \in X$によって生成される主フィルター (principal filter)$\lceil x \rceil$ は，次

のように定義される．

$$\lceil x \rceil \triangleq \{y \mid y \preceq x\}.$$

以下，記号^は，定義であることを示すために使用する．

主フィルター $\lceil x \rceil$ は，x およびその下位概念語の集合であると考えることができる．その要素は‘似ている’と考えることができるから，位相の導入に当っては，主フィルターは開集合となるようにすべきであろう．しかしながら，後で示すように，ある集合 A に主フィルター $\lceil x \rceil$ を対応させる演算は，内包演算とはなりえない．　性質 (v) を基に，A に含まれるすべての主フィルターの和集合を A の開核とするような演算を定義する．すなわち，次のような演算$^\diamond$を定義する．

$$A^\diamond \triangleq \{z \mid z \in \lceil y \rceil \wedge \lceil y \rceil \subseteq A\}.$$

すると，演算$^\diamond$は，内包演算であることが証明される．　$\lceil x \rceil^\diamond = \lceil x \rceil$ であることは，定義から明かであるので，$\lceil x \rceil$ は，開集合である．一方，すべての主フィルターが開集合であるという要請から，内包演算を定義した．そこで任意の内包演算$^\circ$に対し，$\lceil x \rceil^\circ = \lceil x \rceil$ であることを要請したい．すると，すべての主フィルターを開集合とする内包演算は，$^\diamond$以外は存在しないことが証明される（文献3・定理2）．つまり，任意の集合に対して一つの主フィルターを開核として写像する内包演算は存在しない．

このことにより，ある不自然さが生じる．例えば，オントロジーの一つの極小要素からなる集合は，主フィルターであるので，開核であり，開集合である．とこれで，性質 (vi) により極小要素の任意の集合は，開集合である．つまり，ここでの仮定では，極小要素はすべて互いに‘似て’しまうのである．

2．2　開集合と直下語

オントロジーにおいて，語 x が語 yのすぐ下の概念であるというのは，x は yの下位概念であり，x と y の間には，別の概念を表わす語 zが存在しないときであろう．そこで $x \lhd y$と書いて，x が yの**直下**ということを次のように定義する．

$$x \lhd y \overset{\wedge}{\Longleftrightarrow} x \prec y \wedge \neg\exists z(x \prec z \wedge z \prec y), \quad (x \prec y \overset{\wedge}{\Longleftrightarrow} x \preceq y \wedge x \neq y).$$

すると，次の定理が得られる[3)]．

定理 1　$A = \lceil x \rceil \Longleftrightarrow A$ は $x \in A$ であるような最小の開集合である．

定理 2 $x \lhd y \Longleftrightarrow \exists A(\lceil y \rceil - A = \{y\} \wedge A$ は開集合 $\wedge \lceil x \rceil$ は A に包含される極大主フィルター$)$．

3．オントロジー作成手続き

共通の性質をもつ集合を開集合であると仮定する．ある世界についてのすべての開集合が観測されたならば，2.2 節で述べた性質に基づき，その観測からオントロジーをつくることができる．次の手続きは，定理 1 と定理 2 に基づき，ボトムアップ的に直下関係を検知してゆくものである．

＜開集合族からオントロジーをつくる手続き＞

すべての開集合の族をΩと書く．集合 A の要素数を $|A|$ で表わす．$\max_{A\in\Omega}|A| = m$ とする．また，$\mathcal{O}_k(A) \stackrel{\Delta}{\Longleftrightarrow} |A| = k \wedge A \in \Omega$．$Y$を集合の集合とし，$Y$の要素から和集合をとる演算によってつくられるすべて集合の集合をYの和集合閉包といい，$\bigcup^+(Y)$ で表わす．

ステップ1　$X = \emptyset$．$\Upsilon = \emptyset$．$\Gamma = \emptyset$．$k = 1$．

ステップ2　$\mathcal{O}_k(A)$ のとき，$A = \{a\}$ ならば，$X \leftarrow X \cup \{a\}$，$\Upsilon \leftarrow \Upsilon \cup \{\{a\}\}$ とせよ．$\mathcal{O}_k(A)$ であるすべての A について，この手順を繰り返せ．

ステップ3　$k \leftarrow k+1$．$k > m$ ならば，終了せよ．

ステップ4　$\Upsilon \leftarrow \bigcup^+(\Upsilon)$ とせよ．$\mathcal{O}_k(A)$ のとき，ある $B \in \Upsilon$が存在して，$A - B = \{a\}$ ならば，$X \leftarrow X \cup \{a\}$，$\Upsilon \leftarrow \Upsilon \cup A$ とせよ．$b \in B \cap X \wedge \neg\exists x(b \lhd x \notin \Gamma)$ であるすべての b に対して，$\Gamma \leftarrow \Gamma \cup \{b \lhd a\}$ とせよ．$\mathcal{O}_k(A)$ であるすべての A について，この手順を繰り返し，ステップ3に行け．　□

この手続きによって，Ωからオントロジー $(X, \preceq)$ が得られる．関係$\preceq$は，Γによってあたえられる．

4．英文科学技術文からのオントロジー自動作成の試み

オントロジーに位相を導入し，類似した概念の集まりを開集合と仮定し，すべての開集合が観測されたと仮定すると，3節の手続きによりオントロジーを自動作成するすることができる．しかし，自動作成手続きの核心部分は，定理2によるものであり，開集合の差集合がシングルトンかどうかを見る神経質な判定を含んでいる．2.1節で指摘した問題もあり，現実世界のデータに対しては，すべての開集合が観測されると考えるのは非現実的であり，3節の手続きをそのまま適用するわけにはいかない．

そこで，3節の手続きを基に現実的なオントロジー自動作成法を開発するための指針を得るためにINSPECテープの抄録文から，オントロジーをつくることを試みた．方針は次のとおりである．

1) Swallow is swift と Airplane is swift のような文があれば，swallow と airplane は似ているとする．すなわち，$\{x|x$ is/are swift $\}$ を一つの開集合とみなす．そこで，形容詞の個数だけ開集合が存在することになる．

2) Swallow is black and swift などのように形容詞が接続詞で連結されている文では最初の形容詞だけをとり，Swallow is baack とみなす．

3) Swallow is not yellow における not は無視した．not だけではなく，すべての副詞を無視した．

(2) と (3) の方針をとったのは，まだ連言／選言句の範囲決定ができないこと，scarecely などの準否定の副詞の意味をとらえることができていないことなどによる．

対象とした抄録文は，1989年から1993年の5年間の天文学に関する494,253文である．著者らは，動詞決定法を開発している[4]ので，これを使ってbe動詞による114,607の繋辞文を抽出した．このうち，上記のように‘x is/are 形容詞’の形式の文の数は，12,375であり，重複を除くと，9,780であった．これらの文に現れる形容詞の異なり語数は，1,176であり，表1に高頻度の形容詞を示した．

一方，xとなる名詞の異なり語数は，2,617であった．そこで，各形容詞には2,617からなる名詞の集合の部分集合が対応する．これらの部分集合の包含関係を調査したところ，4,626個の包含関係が存在していることがわかった．この関係から開集合の順序構造を求めることができる．

さて，3節の手続きは，基本的には開集合の包含関係に基づいている．そこで，二つの名詞 x と yについ

表 1: 高頻度の形容詞

形容詞	頻度	形容詞	頻度
important	979	negative	105
small	680	dominant	103
present	603	far	98
available	510	flat	96
large	492	true	94
stable	341	positive	94
significant	318	uncertain	88
low	312	unknown	87
valid	286	good	87
unstable	284	complex	86
high	281	observable	85
constant	241	homogeneous	84
evident	206	clear	84
weak	205	uniform	81
still	192	effective	77
variable	188	consistent	76
negligible	181	useful	72
likely	168	simple	72
visible	167	linear	71
absent	150	isotropic	71
strong	141	normal	70
detectable	137	difficult	69
common	133	asymmetric	68
apparent	132	typical	67
accurate	130	symmetric	67
correct	129	thin	65
necessary	126	short	64
active	116	rare	63
complete	108	unusual	62
unique	107	enough	62

表 2: ISA 関係

名詞 ≻	名詞	頻度	名詞 ≻	名詞	頻度
fluid	source	10	zone	fault	2
star	progenitor	5	wind	drag	2
star	one	5	waves	source	2
star	component	5	wave	field	2
technique	method	4	transition	depth	2
step	analysis	4	transit	eclipse	2
star	companion	4	tool	use	2
order	mass	4	tool	radar	2
nova	object	4	tool	method	2
extension	work	4	time	correlation	2
continuation	work	4	technique	use	2
continuation	paper	4	technique	spectroscopy	2
attempt	paper	4	technique	imaging	2
variable	star	3	task	selection	2
technique	radar	3	task	location	2
technique	analysis	3	system	star	2
system	object	3	system	line	2
star	compact	3	superposition	emission	2
spectrum	signature	3	supergiant	progenitor	2
sequel	paper	3	step	study	2
process	star	3	star	primary	2
problem	determination	3	star	object	2
power	spectrum	3	spiral	galaxy	2
parameter	intensity	3	spheroid	primary	2
function	amount	3	spheroid	body	2
extension	paper	3	source	lack	2
equation	result	3	source	image	2
dwarf	star	3	scintillation	purpose	2
candidate	neutralino	3	scalar	phi	2
attempt	study	3	scalar	effective	2

absorption ≻ emission ≻ component ≻ one ≻ energy ≻ radiation ≻ fluid ≻ flow ≻ current ≻ mechanism ≻ fluctuations ≻ noise ≻ ocean ≻ model ≻ approximation ≻ gas ≻ source ≻ areas

absorption ≻ emission ≻ component ≻ one ≻ energy ≻ radiation ≻ fluid ≻ flow ≻ current ≻ mechanism ≻ fluctuations ≻ noise ≻ ocean ≻ model ≻ approximation ≻ gas ≻ source ≻ boundary ≻ estimate

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ energy ≻ accretion

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ energy ≻ conversion ≻ (component)

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ (one)

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ plasma ≻ effect ≻ anomaly

absorption ≻ emission ≻ component ≻ one ≻ energy ≻ accretion

absorption ≻ emission ≻ component ≻ one ≻ energy ≻ conversion ≻ (component)

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ plasma ≻ energy ≻ extinction

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ plasma ≻ energy ≻ (radiation)

absorption ≻ emission ≻ component ≻ one ≻ model ≻ approximation ≻ method ≻ technique ≻ analysis ≻ radiation ≻ fluid ≻ source ≻ time ≻ distribution ≻ wind ≻ explanation ≻ mechanism ≻ wave ≻ field ≻ plasma ≻ energy ≻ result ≻ (one)

図 1: 名詞の階層構造

て，$x \prec y$あるいは$y \succ x$と書いて，yはxの上位概念であることを次のように定義しても，2節と矛盾は生じない．

$$x \prec y \overset{\wedge}{\Longleftrightarrow} \exists A \exists B(x \in A \wedge y \in B \wedge A \subset B \wedge x \notin B)$$

ここで，AとBは開集合．

$\prec$ ($\succ$) 関係を自動的に検証するために，上述の114,607の繋辞文からISA関係を抽出した．抽出法は，名詞修飾語や副詞を無視し，‘名詞1 is a/an 名詞2’の形の文があれば，名詞1と名詞2にはISA関係があるとした．名詞1と名詞2は被修飾名詞であり，この決定には著者らが開発した方法[5]を用いた．連言／選言句の処理は形容詞の場合と同じである．抽出処理を行ったところ，3,888の繋辞文が上記の形をしていて，重複を除くと3,674のISA関係が得られた．

$\prec$関係があり，かつISA関係があるものは，543組であった．また，この関係に現れる異なり名詞数は353であり，いずれも1/10程度に減少した．これらの関係の一部を表2に示してる．この関係を基にオントロジーをつくることができる．ここでは，図1に$\succ$系列の一部を示した．系列の最後に括弧で囲まれた名詞が出現しているのは，既に出現した名詞が再び出現したことを示す．すなわち，サイクルが生じたことを意味している．図1において比較的長い系列が存在することは，注目に値する．

5．むすび

本稿では，科学技術文からオントロジーを自動作成する試みについて述べた．自動作成法としては，実用に耐える段階にはないが，各種試行を行い，改良を加え，実用化を目指したいと考えている．本稿の実験は，そのための第1歩である．

なお，本研究は一部，平成7年度科研費補助金（一般研究(A)07408007）の補助を受けた．

参　考　文　献

1) Dahlgren, K. and McDowell, J. : Knowledge Representation for Commonsense Reasoning with Text, Computaional Linguistics, **15**, 3, 149-170 (1989).

2) Rosch, E., Mervis,C.,Gray, W. D., Johnson, D. M., and Boyes-Braem, P.: Basic Objects in Natural Categories, Cognitive Psychology, **8**, 382-439 (1976).

3) Matsuo, F. : Topology on Ontology, Memoirs of the Fuculty of Eng., Kyushu Univ., **54**, 1, 25-29 (1994).

4) 竹田正幸, 松尾文碩：英文科学技術抄録文における名詞の決定, 情報処理学会論文誌, **34**, 9, 1931-1936 (1993).

5) 竹田正幸, 須田淳一郎, 楠本曲孝, 松尾文碩：英文科学技術抄録文における名詞の決定, 情報処理学会論文誌, **36**, 8 (印刷中).