人工知能学会研究会資料
SIG-SWO-A602-04

自然言語処理用の上位オントロジーと大規模オントロジーの試作

The Trial Development of Upper and Large-Scale Ontologies for Natural Language Processing

荒川直哉
ARAKAWA Naoya

株式会社ジャストシステム
JustSystems Corporation

自然言語の意味処理のために汎用(ドメインに依存しない)オントロジーを構築する試みについて概要を述べる。試作している汎用オントロジーは2つの部分に分けられる。一方は上位オントロジーであり、IEEE の SUMO(Suggested Upper Merged Ontology)を改変したものである。他方は大規模汎用オントロジーであり、EDR コーパス中に出現する概念識別子を用いたオントロジーである。前者の概念数は現在 977 個であり、後者の概念数は5万件余りである。双方とも OWL-DL で記述されている。

1. はじめに

著者らは自然言語処理にオントロジーを用いることを目指している。オントロジーは広い範囲の自然言語処理に用いることができる。オントロジーはルールに基づく処理や統計ベースの処理におけるタームの汎化(あるいはスパースネスの解消)に役立つ。具体的には、係り受け解析や照応解析などの解析系の処理や統計ベースの検索にオントロジーを用いることができる。一方、オントロジー本来の力をより発揮させることができる分野は意味解析である。ここで意味解析とは、自然言語のテキストから(できれば曖昧性のない)論理的な表現(意味表現)を得る変換処理を指す。自然言語の意味表現はオントロジーの語彙によって構成される。テキストからの意味表現を集積することにより知識ベースを構築することができるが、知識ベースへのクエリに際してオントロジーに基づいた推論を用いると、クエリを知的なものとすることができる。

このように自然言語処理におけるオントロジーのメリットは明らかであるものの、従来日本語の自然言語処理に用いることができる汎用オントロジーは一般には入手することができない(英語用には WordNet [1] や OpenCyc [2] といったリソースが存在する)。そこで、著者らは日本語の自然言語処理に用いることができる汎用オントロジーを構築する試みを開始した。汎用オントロジーとしては、諸概念の整理の基準となる上位オントロジーと、日本語の諸概念をカバーする大規模オントロジーの両者が必要であると考えられた。以下、前者を同グループの通称「Gnosis」を用いて GUO(Gnosis Upper Ontology)、後者を EDRG と仮称し、これらのオントロジーの構築の試みについて述べる。また、日本語の語彙をオントロジーの語彙に結びつける辞書にも言及する。

連絡先: 荒川直哉, (株)ジャストシステム,
東京都港区北青山1−2−3 青山ビルヂング
naoya_arakawa@justsystem.co.jp

2. 上位オントロジー

2.1 SUMO について

著者らは、上位オントロジー[3]の候補を検討し、コンパクトで比較的に理解しやすいように思われる SUMO[4]を第一候補として選択した。著者らはセマンティックウェブ技術[5]を用いることを考えていたため、オントロジーは OWL[6]で記述されていることが望ましかった。SUMO は元来 KIF で記述されているが、SUMO を管理していた Teknowledge 社[7]は OWL 版も作成していた。著者らはこの OWL 版 SUMO の検討を開始したが、そのまま用いるには多くの問題があることが判明した。以下に問題点の一部を記す。

- KIF 版 SUMO にはルールによる公理が記述されているが、OWL 版ではそれらの公理は削除されている(OWL オントロジーとして用いるならこれは問題ない)。
- OWL の語彙を用いずに独自の語彙を用いている(disjoint など)。
- OWL の語彙に変換できると思われるが、単に削除されている関係記述がある(partition、exhausitiveAttribute vs. owl:distinctMembers など)。
- 項が3以上の述語は無視されている。
- OWL-Full であり、推論に用いるには不適切である。

次に概念階層についても検討を行ったが、多くの点で違和感を覚えた。その理由は、SUMO の概念ラベル(ID)と定義内容がしばしば乖離していることである。例えば、Substance は分割してもその性質が変わらないものと定義されているが、その下位概念として原子や分子が挙げられている。また、Substance と排他的な概念として CorpuscularObject(「構造物体」)があり、その下位概念に Artifact(人工物)や ContentBearingObject (いわゆるメディア)が存在している。CorpuscularObject は SelfConnectedObject の下位概念だが、すべての人工物やメディアが

SelfConnectedObject ではないと思われる(例えばソフトウェアや電子メディア)。もちろん、Artifact や ContentBearingObject を SelfConnectedObject であるような人工物やメディアと定義することは自由だが、人間がアノテーションなどの目的でオントロジーを使用する場合、定義とラベルの乖離は混乱を招くため、できるだけ避けることが望ましい。抽象的なものに関しても、学問を Proposition の下に置いたり(学問は単なる命題の寄せ集めではなく実践を含む)、映画が Text(言語表現の一種として定義されている)の下に分類されたりしており、見直しを行う必要があると思われた。

こうした課題にもかかわらず、SUMO は大半において再利用可能と思われたため、著者らは SUMO をベースとして新しい上位オントロジーを作成することにした。

2.2 GUO

SUMO の問題点を修正し、OWL-DL の意味論に沿った形に再構成したものが GUO である。現在の GUO を SUMO と比較すると、SUMO の概念のうちおよそ 360 件が削除され、400 あまりの新規概念が導入されている(これらの数は概念ラベルの変更も含んでいる)。すなわち、40%程度の概念に関して入れ替えまたはラベルの付け替えが行われたことになる。当初、最上位概念階層の整理が行われ、次に、後述する EDR 概念識別子とのマッピング作業において必要とされた概念が追加された。また、自然言語の意味表現の設計に従い、意味表現に必要な概念も追加された。削除された概念には、自然言語応用では通常使用されないと考えられる数学的概念などが含まれる。表1に OWL の概念分類に従った概念数の内訳を示す。個体の多くは(自然言語の意味を表現する)抽象的な属性値を表す。

種別	件数
クラス	605
個体	55
ObjectProperty	310
DatatypeProperty	7

表 1

(1) クラス階層

GUO は、SUMO の最上位階層に見られる Physical/Abstract の2分法の代わりに、やや雑多に見えるいくつかの概念(例えば Spatial や Temporal、Agent など)を持ち込んでいる。GUO の最上位階層はしたがって EDR により近いものになっているが、これは自然言語の意味を表記するという共通の目的のために類似の構造を選択したということになる。現在の GUO の最上位構造には後述するように議論もあり、完全にフィックスされたものではない。

現在の GUO の最上位クラス階層は図 1のようになっている。

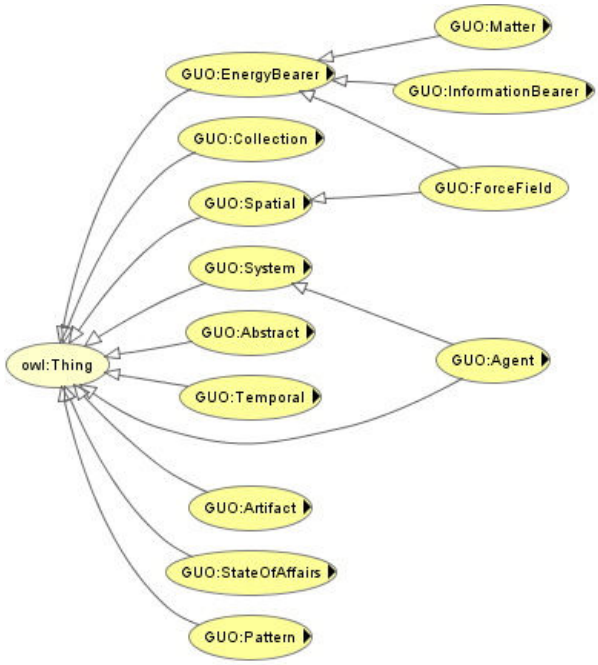

図 1

・Collection (集合)
ものごと(owl:Thing)は Collection とそれ以外(個体)に分けられる。SUMO:Collection は Physical なものだけだが、ここでは SUMO:SetOrClass を含め、複数要素の集まりをすべて Collection とする。なお、デフォルトは個体であって、特に個体クラスというものは設けていない。
・Abstract (抽象的・エネルギーの媒体とみなされないもの)
・EnergyBearer (エネルギーの媒体とみなされるもの)
素粒子、物理力場、物質一般、波動を含む。現代物理学で説明されるものとも限らず、魔術的なものや霊的なものを含む(従って GUO は現実世界だけを記述対象にしているわけではない)。
・Spatial (空間的なもの)
・Temporal (時間的なもの)
・Agent (有意志でありうるもの)
現状、EnergyBearer を継承していない(たとえば「神」が EnergyBearer でないと考える人もいるかもしれない)。
・StateOfAffairs (事態)
・System(システム)
相互に関係する複数の要素を1つのエンティティとしてみるもの
・Artifact (人工物)
多くの場合 EnergyBearer だが、計算機言語やソフトウェアは Artifact であっても EnergyBearer ではない。
・Pattern(パターン)
時空間上の強弱変化を持つもの

Abstract（抽象的なもの）の下位概念は次のようになっている。
・Attribute（属性値のクラス）
・IdeationalContent（思考内容・SUMO の Proposition に相当）
Procedure（手続き）、Rule（法則）、IdeationalSituation（表象された状況）を含む。
・Language（言語）⊂System
・Institution（制度）⊂System
・Extent（はじまりと終わりの値を持つ範囲・TimeInterval を含む。）
・MathematicalObject（数学概念・Relation, Number, Graph を含む。）

Abstract の下位概念中で Attribute（図 2）は特に重要なクラスであり、ほとんどの形容詞に由来する概念や量の概念が Attribute の下位概念である（もっとも個々の形容詞概念はほとんど GUO には収録されていない）。現在、（SUMO と同様に）職業などの役割概念も Attribute の下位概念とされている。また、（SUMO とは異なり）金属や固体など物質の種類や状態を示すカテゴリも Attribute の下位に置かれる。なお、虚構的な実体（《シャーロック・ホームズ》など）は、Attribute のインスタンス Fictitious を属性値として与えることによって実在のものと区別する。

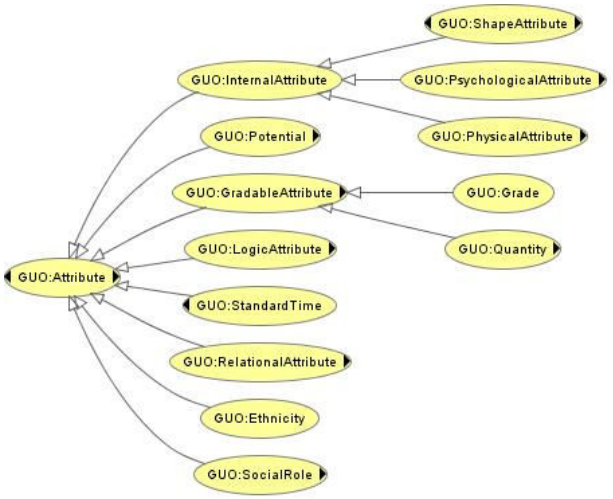

図 2

StateOfAffairs（事態）は自然言語の動詞が表現する概念に相当する。これに近い SUMO の概念は Process（時間的な部分を持つ生起するもの）であるが、瞬間的な事象や静的な事象（「持っている」など）を含むのかどうかはっきりしない。StateOfAffairs は下位概念として Event（動的な事態）と State（静的な事態）を持つ。SUMO の Process 概念は、Beth Levin[8]の動詞分類を参考にしているとされる。GUO の Event 分類もほぼそれを踏襲しているが、語彙概念構造（Lexical Conceptual Structure）理論[9][10]を参照して見直しを行い、Event の下に Change や CausalProcess といった基本的な概念を追加した（図 3）。

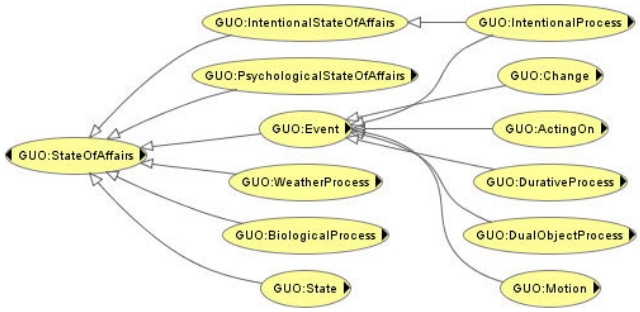

図 3

Event の下には他に DurativeProcess（一定期間続けることができる Event）、IntentionalProcess（意図的な Event）、Motion（動きを伴う Event）、DualObjectProcess（片方が必ずしも原因となっていないような2つ以上のオブジェクトに関わる Event）、BiologicalProcess（生物的・生理的な Event）、WeatherProcess（気象現象）などが含まれる。

(2) GUO のプロパティ

ここでプロパティとは（OWL の用語法に従い）2項関係を指す。GUO は、SUMO 由来のプロパティに加え、自然言語の意味表現に用いられる様々なプロパティを持つ。その一部を以下に示す。
・logicallyRelated（論理的関係）
precondition, consequent, inconsistentWith, entails など。談話関係（逆接、順接など）もここに含まれる。談話関係の設定においては SDRT 理論[11]を参考にした。
・temporallyRelated（時間的関係）
timePosition に関する様々な関係を含む。
・spatiallyRelated（空間的関係）
above, below などの位置関係の他、部分関係を示す part もここに属している。
・semanticRole（意味役割）
StateOfAffairs を定義域に持ち、StateOfAffairs インスタンスに関与するオブジェクトを関連づける（SUMO の CaseRole に相当）。上記の語彙概念構造理論、EDR/UNL[12]の関係子を参考にして調整を行った。actor, origin, method など。
・causallyRelated（因果関係）
causes, prevents, output, inhibits を含む effects の他、input, hasPotential, reason を含む。
・attribute（属性）
オブジェクトと Attribute クラスのインスタンスを結びつけるプロパティ。量を表す quantity、程度を表す hasExtent、真理値を表す truth、比率の分母を表す per を含む。逆関係の attributeHolder は、属性クラス Attribute のサブクラスがどのようなものごとの属性なのかを指定する目的で用いられる。

(3) デザインポリシーと課題

現在の GUO は、武田[3]や溝口[13]で言及されているいくつかの文献を参考にしているものの、強いデザインポリシーに準拠して作られたオントロジーではない。強いポリシーを選択する論拠にコミットしていない（できていない）ため、GUO は中庸的なものにな

っている。GUO は当面使用できるオントロジーとして発展中のものであり、現在のデザインは仮のものである。今後、さまざまなオントロジー構築方法論[14]の検討による改良が想定される。以下、デザイン上のいくつかの留意点を述べる。

・最上位階層
現在の最上位階層が妥当なものかどうかについては議論がある。例えば、Artifact 概念は Attribute の下位概念としたほうがよいのかもしれない。また、Temporal や Spatial は(おそらくは Agent も)カテゴリというより整理のためのラベルにすぎない。

・ロール(役割)概念
GUO では、ロールはロールを表す属性クラスと attribute プロパティの組み合わせによって表現される。実際の意味表現では、例えば1人の教師は「attribute プロパティの値が《教師》ロールクラスのインスタンス(抽象物)であるようなもの」と表現される。

・多重継承
ものごとには唯一の本質があるとする存在論においては多重継承が忌避されるかもしれない。しかし現在の GUO では本質に関する議論を保留しており、多重継承を許容している。例えば《人間》は Agent と EnergyBearer を継承し、《パソコン》も System と EnergyBearer、Artifact を継承する。また、Agent は System を継承している(図 1)。

・公理の密度
GUO は軽量オントロジーであって、全概念数(977 個)に対する公理の個数は比較的少ない(2430 件:1 概念あたり 2.49 個、内訳は表 2 参照)。これは GUO の作成目的が自然言語の意味表現で用いられる概念を整理するためであることを反映している。

GUO 内の OWL 公理	個数
rdfs:subClassOf	1149
rdfs:subPropertyOf	556
rdf:type (個体)	55
owl:FunctionalProperty	248
owl:SymmetricProperty	59
owl:TransitiveProperty	52
rdfs:range	179
rdfs:domain	113
owl:hasValue	4
owl:differentFrom	7
owl:disjointWith	8
その他	0
計	2430

表 2

なお、GUO のほとんどの概念はプリミティブである。すなわち、(必要)十分条件が与えられていない。これは、GUO の概念を用いて定義できる概念はそもそも GUO 内で定義する**必要**がないという考えによるものである。現在、利便性のために定義済みの概念も若干採録しているが、ファイルを分けて管理することも考えている。

(4) その他のトピック

・TimePosition(時間上の位置)
SUMO では Quantity の下位概念になっているが、時間上の位置は《量(quantity)》であるとは言いがたいため、独立したカテゴリとして Temporal の直下に置いた。

・Process(イベント列)
上記のように SUMO の Process は、GUO の Event に相当する。GUO の Process はイベント列であり、下位概念として Activity、Cycle、Procedure を含む。Procedure は Abstract も継承するので、Process は実際の Event 列と抽象的な Event 列を包含する汎化クラスとなっている(こうしたクラスの作り方には問題があるかもしれない)。

・記号論
本などのメディア(ContentMedium)とその内容(IdeationalContent)は区別して扱う。ContentMedium 周辺の構造は図 4のようになっている。

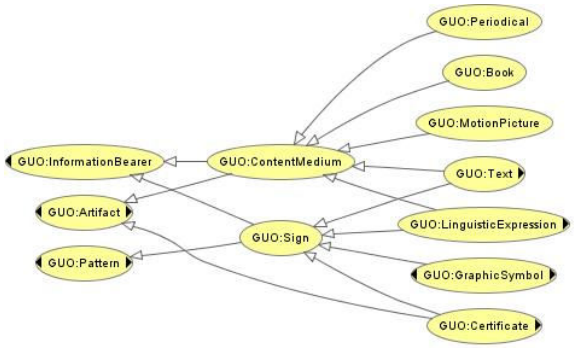

図 4

・単位
SUMO では、単位を用いた量は2項**関数** MagnitudeFn(数, 単位)を用いて表現する。これは全体として3項関係となるため RDF/OWL ベースでそのままは利用できない。現在、単位表現についてオントロジーを調整中である。

・場所的関係およびメレオロジー
SUMO では「上」などの位置関係は属性クラスによって表現されるが、GUO ではプロパティ(2項関係)として表現される。メレオロジー(部分全体論)は SUMO の体系をほぼ継承しているが、《前側(anteriorPart)》など、どの部分を指すかを表すプロパティなどを追加している。

・物体
SUMO の項で問題とした Substance と CorpuscularObject は、GUO では UniformBody と StructuralObject と改名されている。これら概念の近傍を図 5に示す。

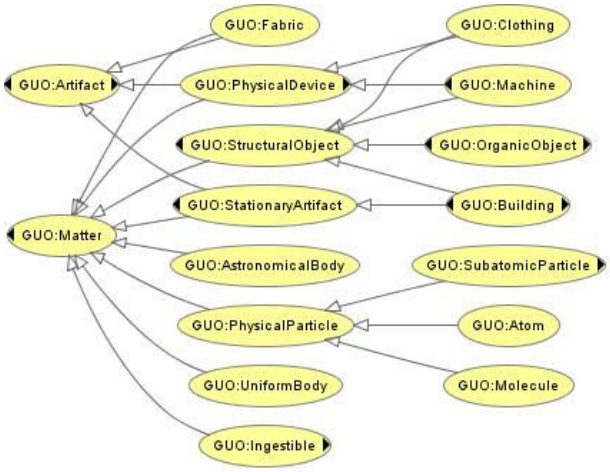

図 5

3. 大規模汎用オントロジー

EDR[15]の概念体系をベースにして、日本語の諸概念を含む実験用オントロジー(EDRG)を作成している。実際に作業対象となっているのは、EDR コーパスに出現する概念とそれらの上位概念(計約6万概念)である。現在までに、それらの概念と GUO 概念のマッピングを終了している。また、(EDR 概念体系では同義概念の集約が行われていなかったため)同義概念の集約作業(約 6,900 件)も行った(作業後の OWL 化概念数は約 52,000 件)。現在の概念内訳は表 3 および表 4 を参照されたい。

クラス 総数	41,535
プロパティ 総数	4,308
個体 総数	5,933
計	51,776

表 3 EDRG の概念内訳

GUO上位概念	クラス	個体
Abstract	9262	4090
└Attribute	└ 7149	└3962
Collection	1778	393
EnergyBearer	9263	97
Spatial	4327	1494
Temporal	3697	325
Agent	2390	1271
StateOfAffairs	16012	68
System	7062	1324
Artifact	6522	89
Pattern	4654	205

表 4 EDRG の上位概念別概念内訳

GUO へのマッピングの目的は次の通り。

・意味解析においては GUO の概念を用いた処理を行いたい。EDR の概念を直接用いると、処理全体が EDR の体系に依存してしまう。

・マッピングによって EDR 概念体系の不都合な点を改変することができる(下記)。

また、この作業の副次的な効果として GUO を洗練することもできた。データ作業は表形式で行い、ツールを用いて OWL 形式に変換している。

概念の GUO のマッピングにおいては、次のことに留意した。

・上位下位関係の厳密性
概念Aによって表される個体の集合が概念Bによって表される個体の集合をすべて含むとき、概念Aは概念Bの上位概念である。EDR の概念体系はこの水準の厳密性を持っていないが、意味表現用のオントロジーとしては厳密さが要求される。データの OWL 化の際には、EDR の概念階層を GUO へのマッピングと矛盾しない部分のみ OWL にインポートすることにより上位下位関係の厳密性を保証している。なお、EDR の概念階層の妥当性について予備調査を行い、さまざまな問題があることが判明した。マッピングによる矛盾リンクの排除のみでは適切な階層を構築することにはならない。階層構造の再構築は今後の課題である。

・個体の設定
EDR の概念体系はクラスと個体を区別していない。(主に固有名詞で表現されるような)唯一存在するものは個体として登録される。

・プロパティの設定
EDR の概念体系はプロパティを区別しないが、EDRG では区別を行う。例えば、《父親》概念は GUO プロパティ father に写像される。また、《遺失物》のような動詞に相対して定義される概念は、Losing という Event 概念を定義域に持つ目的格プロパティ theme の下位プロパティに写像される。

・派生的概念
例えば《先生》という抽象的な役割概念と、その役割を持つ人という概念は派生的関係を持つ。また、《長さ》概念は《長い》という概念と派生関係を持つ。一定の派生的関係を持つ複数の概念は片方のみを使用するような方針をとっている。(EDRG の使用者は、こうした概念の使用に際して文脈に応じた自動的な派生を行う必要がある。)

EDRG は現在以下のような情報を持つ。

・概念区分
OWL の仕様に沿って、クラス概念、個体概念、プロパティ(2項関係・ObjectProperty と DatatypeProperty)概念を区別している。

・概念間の上下関係(上記)

・ある概念が表すものが常に持つ属性
たとえば《王女》概念が表すものは常に Female という属性を持つ。

・プロパティ(2項関係)概念の定義域と値域
プロパティ概念に対してできるかぎり定義域と値域

を設定している。たとえば(通常の)親子関係の定義域と値域はともに Agent である。

・属性概念の付与対象指定
現在開発中のオントロジーでは、属性概念はクラス概念の一種として表現される。属性概念に対しては、その属性が付与される対象のクラスを指定することができる。たとえば《寄せ棟造り》という属性の付与対象は《屋根》というクラスに属する。

・集合とその要素
集合を表す概念に対して、その要素を指定することができる。たとえば、《諸国》という概念の要素(メンバー)は《国家》である。

4. オントロジー連携辞書

オントロジーは、単体では自然言語処理に用いることはできない。最大の理由は、オントロジーの概念 ID は任意の文字列であって、対象となる言語のターム(単語など)の形と何らかの一致があることは期待できないからである。実際、GUO の概念 ID は英単語(列)であり、EDR の概念識別子は6桁の 16 進整数であって、日本語のタームとは一致しない。このため、自然言語のタームに対し、オントロジー概念 ID を関係付けるため、何らかの辞書が必要になる。ここでは、そうした辞書をオントロジー連携辞書と呼ぶ。

上記の EDRG については、オントロジー連携辞書として EDR の日本語辞書をほぼそのまま利用することができる。ただし、EDR の日本語辞書は必要以上の多義を採録しているなどの問題があるため、現在対応を考慮中である。

5. 汎用オントロジーの応用

ここで紹介したオントロジーは、はじめに述べたような、さまざまな自然言語処理応用を念頭に構築されている。著者らは、オントロジーの構築と同時に意味表現の設計を行っている(オントロジーには意味処理設計上の課題をフィードバックする形で修正が行われている)。また、実装面では、テキストを解析して、RDF の意味表現を生成し、知識ベースに保存、さらに自然言語からクエリを生成して知識ベースにアクセスする実験も行っている。これらの意味表現や意味解析の詳細の紹介は別の機会に譲ることとする。

参考文献

[1] WordNet ホームページ:http://wordnet.princeton.edu/
[2] OpenCyc ホームページ:http://opencyc.org/
[3] 武田英明:上位オントロジー, 人工知能学会誌, 19 巻 2 号 pp.172-186 (2004) http://www-kasm.nii.ac.jp/papers/takeda/03/jsai04top-ontology.pdf
[4] SUMO ホームページ (by Adam Pease): http://www.ontologyportal.org/
[5] Antoniou, G. and van Harmelen, F.: CD-ROM で始めるセマンティック Web, ジャストシステム (2005)
[6] W3C OWL ホームページ: http://www.w3.org/2004/OWL/
[7] Teknowledge 社 SUMO ホームページ: http://ontology.teknowledge.com/
[8] Levin, B.: *English Verb Classes and Alternations: A Preliminary Investigation,* Univ. of Chicago (1993)
[9] 影山太郎:動詞意味論－言語と認知の接点, くろしお出版 (1996)
[10]Jackendoff, R.: *Semantic Structures*, MIT Press (1990).
[11]Asher, N. and Lascarides, A.: *Logics of Conversation,* Cambridge Univ. (2005)
[12]UNL ホームページ:http://www.undl.org/
[13]溝口理一郎:オントロジー工学, オーム社 (2005)
[14]上田俊夫:概念化アスペクト:オントロジー構築の手掛かり, 人工知能学会セマンティックウェブとオントロジー研究会資料 SWO-A602-03 (2006)
[15]EDR ホームページ: http://www2.nict.go.jp/r/r312/EDR/J_index.html