1G4-3

概念間の関係に注目した専門文書解析とLOD技術による バイオミメティクス・オントロジーの大規模化の試み

A Trial to Expand a Biomimetics Ontology using Technical Document Analysis and Linked Open Data Technique focusing on Relationships between Concepts

多田 恭平[*1] Kyohei Tada　古崎 晃司[*1] Kouji Kozaki　來村 徳信[*1] Yoshinobu Kitamura　溝口 理一郎[*2] Riichiro Mizoguchi　駒谷 和範[*1] Kazunori Komatani

[*1] 大阪大学産業科学研究所 The Institute of Scientific and Industrial Research (ISIR), Osaka University

[*2] 北陸先端科学技術大学院大学 Japan Advanced Institute of Science and Technology

It is required to develop biomimetics database that supports engineers to develop new products inspired by biological functions. Biomimetics ontologies are the foundation of the biomimetics database. This article proposes methods to expand them using two techniques. The first is an analysis of the biomimetics documents based on natural language processing. The second is a Linked Open Data technique. This article discusses methods to propose concepts and relationships among them that should be added to the biomimetics ontologies.

1. はじめに

近年，生物がもつ優れた機能を模倣し，技術開発やものつくりに生かすことを目指しているバイオミメティクス分野の研究が盛んに行われている．それに伴って，昆虫や魚類，鳥類といった生物が有する多様性に関する知識を，様々な観点から検索可能なバイオミメティック・データベースの構築が求められている[下村 10]．このデータベースにより，新たな技術を開発しようとする工学研究者に対し，技術革新の着想につながる発想支援の実現を目指している．具体的には,例えば，工学研究者が求める「機能」から，その機能を実現している「生物」を検索するといった利用が想定されている．その際に，より発想を刺激するためには，機能と生物の直接的な関係のみならず，機能の達成関係，生物の生態環境と関連する機能など，様々な関係を用いた検索が重要となる．

そのような発想支援のための検索を実現する為に，本研究では，バイオミメティクスに関する様々な知識を体系化した，バイオミメティクス・オントロジーの構築を進めている[古崎 13]．そのオントロジーを対象に，利用者の関心に応じた観点から探索する技術を適用することで，多種多様な生物の知識を，利用者ごとにその意味構造に基づいて検索可能となる[北河 13]．

また，目標のデータベースを構築するためには，多くの概念が存在し，かつ概念どうしに意味のあるつながりがあるオントロジーを構築しなければならない．そのためのオントロジーの拡充には 2 つのアプローチが考えられる．オントロジーの質を重視し，開発者が手動で情報を構造的に記述する方法[鳥村 15]と，オントロジーの概念や関係の量を重視し，多種多様な生物の情報を半自動的に必要な概念や関係を追加する方法である．本研究は後者について先行研究を基に研究を行ってきた[多田 14]．

半自動的なオントロジーの大規模化は，専門文書の解析と Linked Open Data（LOD）の 2 つの手法から行っている．そして，両方とも概念間の関係に注目している．概念間の関係には，「昆虫」と「セミ」といったような上位クラスと下位クラスとの間の関係である is-a 関係と，「生物の特徴的機能」や「機能の根拠となる構造」といった，is-a 関係以外の関係がある．それらの関係に注目してオントロジーの大規模化を行う．

以下，2 章では本研究で提案する，バイオミメティクス・オントロジーを大規模化する手法について述べる．3 章では 2 章で述べた手法を，昆虫に関するバイオミメティクスを対象に試行した結果と，そこから得られた課題について述べる．4 章では，本論文の総括と今後の展望について述べる．

連絡先：多田 恭平，大阪大学産業科学研究所 知識科学研究分野，〒567-0047 大阪府茨木市美穂ヶ丘 8-1，℡:06-6879-8416，tada@ei.sanken.osaka-u.ac.jp

2. 概念間関係に注目したオントロジー大規模化

2.1 目指すオントロジー

本研究で大規模化の結果として得ることを目指しているオントロジーとして，図 1 に手動で構築されたバイオミメティクス・オントロジーの一部を示す．オントロジーは概念と概念間の関係で定義される．バイオミメティクス・オントロジーで用いられる関係の種類は，is-a リンクで表される is-a 関係と，スロットで表されるその他の種類の関係がある．is-a 関係は，その概念がどのクラスに属しているのか，すなわちその概念の上位クラスがなにであるのかを表す．図 1 の例では，「カタツムリ」とその上位クラスである「動物」の is-a 関係で表されている．一方，その他の関係を表すスロットでは，「カタツムリ」の“生態環境”は「湿地」である，といった関係性が定義される．スロットの上部に書かれた“生態環境”などのラベルが関係の種類を表している．これらの関係を用いることで，例えば「防汚」機能をもつ生物として「カタツムリ」を検索結果として出力するといった検索システムが構築できる．しかし，これを開発者の手で構築するのはコストがかかるため，本研究ではこれらの関係の付与を自動化することを目指す．

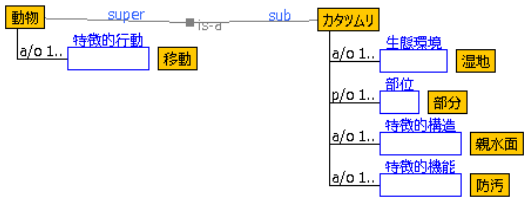

図 1 目指しているオントロジーの形の例[鳥村 15]

2.2 オントロジーの半自動大規模化の手法と手順

バイオミメティクス・オントロジーには，バイオミメティクスのドメインに特化した知識に加えて，より一般性が高く幅の広い知識も，含まれることが望ましい．なぜなら，生物に詳しくない工学研究者を利用対象としているからである．そこで，それぞれの知識について異なる手法から，オントロジー構築の基となる知識を獲得する．ドメイン知識は，バイオミメティクスについて書かれた専門文書を自然言語処理技術を用いて解析することで得る．一方，一般的な知識は，DBpedia など既存の Linked Open Data を利用することで獲得する．

これら2つの手法に共通するオントロジー大規模化の手順は，以下の 3 ステップからなる．

(1)オントロジーに追加する概念の候補の選択

情報源となる専門文書や LOD から，バイオミメティクス・オントロジーに追加する概念の候補となる単語を選択する．

(2)オントロジーに追加する概念の上位クラスの同定

追加候補の単語をオントロジーへ追加する箇所，すなわち，候補概念の上位クラスを同定する．

(3)その他の関係の種類を同定

同定した上位クラスの情報を用いて，追加概念と他の既存概念との間の，is-a 関係以外の関係の種類を同定する．

2.3 バイオミメティクス・上位オントロジー

バイオミメティクス・オントロジーを半自動的に大規模化するにあたり，すべての概念を自動処理で取得することは困難である．そこで，大規模化作業の基となる上位オントロジーを手動であらかじめ構築することとした．これまでに筆者らは，昆虫を対象としたバイオミメティクスの専門書である「昆虫ミメティックス」[下澤 08]に付録されているキーワード索引を用いて，上位オントロジーを構築した．キーワード索引とは，著者が重要であると判断した語をキーワードとして選出し，索引化したものである．

先行研究において，キーワード索引に挙げられている単語 657 語を，19 の上位概念に手動で分類することで，オントロジーを構築した．これをバイオミメティクス・上位オントロジーと呼ぶ[多田 14]．以降ではこの上位オントロジーを基にしてオントロジーの大規模化を行う．

2.4 専門文書解析によるオントロジー大規模化

(1) オントロジーに追加する概念の候補の選択

発想支援のための検索は，オントロジーの概念同士のつながりを探索することで行われる．よって，より良い検索結果を多く得るためには，元のオントロジーに存在する概念と何らかの関係をもつ概念を追加する必要がある．よって本研究では，そのような概念を表す単語を，オントロジーに追加すべき概念候補とする．

具体的には，まず専門文章の本文中にある単語のうち，バイオミメティクス・上位オントロジーに存在する概念と係り受け関係，もしくは同一文内の共起関係にある単語を抽出する．これらの関係の抽出には，テキストマイニング用ソフトウェア「Text Mining Studio[1]」のテキスト文中に出現する単語同士の係り受け頻度，共起頻度を出力する機能を用いる．そして最後に，その単語群から品詞が形容詞語幹名詞，数詞，自立形容詞である単語を取り除く．これは，上位オントロジーの概念にそういった品詞の概念がほとんど存在せず，オントロジーに追加すべきではないと考えたからである．品詞による選別を行ったあとの単語群を追加候補とする．

(2) オントロジーに追加する概念の上位クラスの同定

上位オントロジーの概念と(1)で得た追加候補の単語の類似度を用いて，追加候補の単語の上位クラスを同定する．ここでは，上位オントロジーの概念 U と追加候補の単語 X の類似度が高いとき，それらの概念は，上位クラスが同じ兄弟概念になると考え，U の上位クラスを X の上位クラスとする．単語間の類似度計算には，専門文書の各文における各単語と共起する単語の組を特徴ベクトルとして，自然言語処理分野で一般的によくつかわれているコサイン類似度を用いる[多田 14]．

(3) その他の関係の同定

概念間のその他の関係の同定には，候補選択に用いた係り受け関係と共起関係を用いる．これらの関係が抽出された単語間には，何らかの関連性が認められると考えることができるが，概念間の意味的な関係の種類は区別されていない．そこで，係り受け頻度解析，共起頻度解析によって得られた関連性のある単語の上位クラスの概念を用いて，それらの間の関係の種類を同定する．

同定する関係の種類は，手動で作られたバイオミメティック・オントロジー[古崎 13, 鳥村 15]で定義されている is-a 関係以外の関係のうち，主な関係である 8 種の関係である．このオントロジーでは，関係づけられている 2 つの概念の上位クラスの組み合わせが定義されている．それを利用し，概念間に係り受けか共起による関連性がみられたときに，それらの概念の上位クラスの間の関係の種類を下位に継承させる．

例えば，単語 A と単語 B に関連性が認められ，それぞれの上位概念が「機能」と「生物種」であった場合，「機能」と「生物種」との間に「生物種の特徴的機能」という関係が定義されているので，単語 A と単語 B の間にも「生物種の特徴的機能」という関係名を継承できる．

2.5 Linked Open Data を利用したオントロジー大規模化

(1) オントロジーに追加する概念の候補の選択

情報源とする LOD から 2.2 節で構築したバイオミメティクス・上位オントロジーの概念と，概念名が一致するデータが存在するものを抽出し，それらのデータと関連性のあるデータをオントロジーに追加する概念の候補とする．データの抽出には，SPARQL エンドポイント(LOD 用の検索 API)が公開されている任意の LOD に対する検索ソフトウェア「簡易 SPARQL ツール[2]」を利用する．

(2) オントロジーに追加する概念の上位クラスの同定

情報源とする LOD は，一般的な百科事典として Web 上で構築されている Wikipedia の知識を基に構築された DBpedia Japanese[3]を用いる．DBpedia Japanese には Wikipedia の infobox に記載されている情報を基に構造化された知識をプロパティとしてもつ．生物種などはその上位クラスである type がプロパティとして記述されているためそのまま用いることができる．また，type プロパティが記述されていないものに関しては，そのデータの種別を定義している subject プロパティを上位クラスの同定に用いることができる．しかし，キーワードやキーフレーズも subject プロパティとして記述されるため，subject プロパティの中から種別についての記述を選択する必要がある．

[1] http://www.msi.co.jp/tmstudio/

[2] http://sourceforge.jp/projects/easylod/wiki/EasySPARQL

[3] http://ja.dbpedia.org/

表 1 コサイン類似度による上位クラス同定の割合の一覧

コサイン類似度	組の総数	判定数	同定されていると判断した組の数	同定されていないと判断した組の数	同定できる割合
0.95-1.00	4	4	0	4	-
0.90-0.95	9	9	5	4	55.6%
0.85-0.90	12	12	2	10	16.7%
0.80-0.85	14	14	3	11	21.4%
0.75-0.80	16	16	7	9	43.8%
0.70-0.75	29	29	4	25	13.8%
0.65-0.70	84	44	11	33	25.0%
0.60-0.65	227	39	9	30	23.1%
0.55-0.60	900	51	5	46	9.8%
0.50-0.55	3130	58	8	50	13.8%

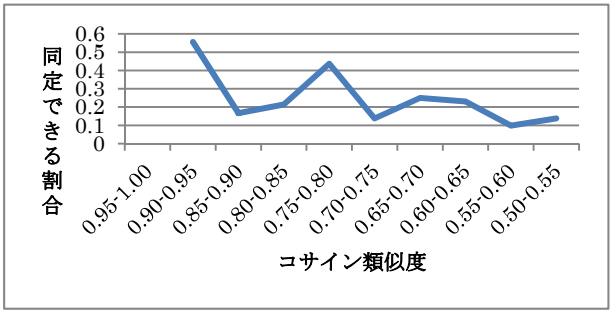

図 2 コサイン類似度の変化における上位クラス同定の割合

(3) その他の関係の同定

上位クラスの同定ができれば，前節で述べた専門文書解析でのその他の関係の同定と同様に，上位クラス同士の関係名を下位概念同士に継承することで同定できる．

3. オントロジー大規模化の試行

3.1 専門文書解析による大規模化の結果・考察

(1) オントロジーに追加する概念の候補の選択

「昆虫ミメティックス」の全本文 16,743 文ついて，「Text Mining Studio」の係り受け頻度解析機能と，ことばネットワーク機能を用いて係り受け関係または共起関係にある単語の組を抽出した．その結果，係り受け関係にある組は 24,378 単語 69,743 組，共起関係にある組は 33,110 単語 254,645 組であることが分かった．

次に，これらの単語の組から上位オントロジーと関連性のある単語を抽出した．まず係り受け頻度解析について，係り元の単語，係り先の単語，およびそれらの両方が上位オントロジーの概念である組の数は，それぞれ 9,223 組， 2,856 組，510 組であった．以上から，係り元単語，係り先単語のどちらかが上位オントロジーの概念である 11,569 組がオントロジーに追加する候補となる関係である．また，その関係を構築している単語の数が 4,182 単語であったため，これがオントロジーに追加する候補となる単語となる．次に，共起頻度解析について，前提単語，結論単語，およびその両方が上位オントロジーの概念である組の数は，それぞれ 780 組，24,408 組，および 243 組であった．以上から，前提単語，結論単語のどちらかが上位オントロジーの概念である 24,945 組がオントロジーに追加する候補となる関係である．また，その関係を構築している単語の数が 14,470 単語であったため，これがオントロジーに追加する候補となる単語となる．そして，得られた 2 つの単語群をマージした結果，17,091 単語が係り受け頻度解析，共起頻度解析から追加候補とできた．

表 2 主な関係名とその関係にある上位クラスの組

関係名	上位クラス A	上位クラス B
特徴的機能	生物種	機能
特徴的構造	生物種	器官，構造
特徴的行動	生物種	生体行動
根拠となる構造	機能	構造
部位	生物種	器官
部分構造	器官	器官
	構造	構造
関連機能	生物種，性質，生体行動，機能	機能
生態環境	生物種	生態環境

最後に，追加候補の単語の品詞を検討した．バイオミメティクス・オントロジーには数詞や助詞といった単語に相当する概念は定義されていないため，除外する必要がある．「Text Mining Studio」の品詞解析機能を用い，サ変接続名詞，一般名詞，形容詞動詞語幹名詞，固有名詞，自立動詞を追加候補とした．その結果，最終的に 14,298 単語をバイオミメティクス・オントロジーに追加する候補として選択した．

(2) オントロジーに追加する概念の上位クラスの同定

「Text Mining Studio」によって得られた全単語のうち，出現頻度が 2 以上である 9,189 単語について，「昆虫ミメティックス」の文中に共起して出現する単語の組 2,134,511 組を生成した．そして，その共起情報を基に単語毎に共起して出現する語の一覧を生成した．これを単語毎の特徴ベクトルとし，これを用いて各単語間のコサイン類似度を計算した．本研究では，類似度が比較的高い 0.5 以上の組で，かつ上位オントロジーの概念と追加候補の単語の組である 4,425 組について，上位クラスが正しく同定できているかを筆者が判定した．

判定結果を表 1 と図 2 に示す．これによると，コサイン類似度により数値にばらつきがあるが，数値が高いものでおよそ半分の割合で自動的に正しく上位クラスを同定できることが分かった．また，追加候補の単語には，「最初」といった一般性の高い単語や，「改定」といった生物に直接関係ない単語といった，追加するのが適当ではないものが含まれ，それが上位クラスが同定できる割合を小さくしているため，追加候補をより洗練することで同定できる精度が向上されることが示唆される．

(3) その他の関係の同定

手動で作られたオントロジー[古崎 13，鳥村 15]における is-a 関係以外の関係を調べると，多く使われている主要な関係が存在することが分かった．オントロジーに存在する 485 個の関係のうち，384 個が 8 種類で構築されていた．

表 2 は，それら 8 種類の関係にある2つの概念の上位クラスを示す．例えば，ある「生物種」とある「機能」に関連性がみられたときに，「特徴的機能」という関係で関連性を定義している．これが 2 つの上位クラスに属する概念において常に成り立つことであるかを調べた．具体的には，バイオミメティクス・上位オントロジーの概念のうち，関連性のある 2 概念がその上位クラス同士をみることで関係名を同定できるかどうかを調べた．

追加する概念の候補の選択に用いた係り受け頻度解析の単語の組 69,743 組と共起頻度解析の単語の組 254,645 組について，組の両単語ともが上位オントロジーの概念であったのは 670 組であった．そのうち，その両単語の上位クラスの組が表 2

表 3　上位クラス同士の関係が下位概念に継承できる割合

関係名	係り受け，共起関係にある組の数（α）	関係名を継承できると判断した数（β）	β／α
特徴的機能	68	64	94.1%
特徴的構造	3	2	66.7%
特徴的行動	36	36	100%
根拠となる構造	3	0	0%
部位	15	15	100%
部分構造	18	13	72.2%
ALL	143	130	90.9%

の組み合わせであるものについて，その関係名が下位クラスに継承できるかを検証した．なお，8 種類の関係のうち，「関連機能」は関係名の定義が曖昧なため，また「生態環境」は上位オントロジーの上位クラスに生態環境が存在しないため今回は検証していない．

表 3 に検証結果を示した．この結果より，提案した手法で，関連性のみられる単語の組の関係名の付与がおおよそ 70%から 100%という高い確率で妥当であったことがわかる．すなわち，本研究で提案した関係の種類の同定方法が，オントロジーの大規模化に有用であることが示唆された．

3.2 LOD 技術による大規模化の結果・考察

(1) オントロジーに追加する概念の候補の選択

「簡易 SPARQL ツール」によって，上位オントロジーの概念 657 語のうち，DBpedia Japanese にデータが存在するものは 287 個（43.7%）であった（2015/03/18 時点）．これらに関して，Wikipedia 上の各記事内でハイパーリンクされている他の見出し語を表す WikiLink という関係が設定されている単語を抽出することで，オントロジーに追加する概念の候補として選択できる．

(2) オントロジーに追加する概念の上位クラスの同定

LOD 技術を用いた上位クラスの同定ができるかを検証するために前節の専門文書解析での追加候補の単語 14,928 語を用いた．「簡易 SPARQL ツール」によって，追加候補単語のうち，DBpedia Japanese にデータが存在するものは 2,386 個（16.0 %）であることが分かった．これについて，DBpedia Japanese において上位クラスが定義されているかどうかを調べた．2,380 個の単語のうち，ランダムに 100 個を選択し，それらの上位クラスが定義されているか調べたところ，7 個に type，50 個に subject というプロパティ（関係）が存在した．このことから，DBpedia Japanese に存在する単語のうちおよそ半分は自動的に上位クラスを同定できることが分かった．しかし，既に上位クラスが定義されているものの中には，複数の上位クラスの下位に分類されている単語が存在する．それらのなかからオントロジーに追加する際の適切な上位クラスを選択することが今後の課題となる．

(3) その他の関係の同定

専門文書解析によるその他の関係の同定は，上位オントロジーに上位クラスとして存在する 19 の概念の間の関係のみを考えればよいため，関係の種類が同定しやすいが，DBpedia Japanese から得られた上位クラスは，上記の 19 概念ではないものも含まれるため，関係名を一意に決めることが困難となる．この課題については，今後解決する必要がある．

4. 本研究の総括と今後の展望

本研究では，生物の多様性に学ぶものつくりを目指したバイオミメティクス分野において，工学研究者に対して新たな製品開発の発想支援ができるようなバイオミメティクス・データベース，またその中核となるバイオミメティクス・オントロジーの大規模化について議論した．

今後の研究に関して，専門文書解析での大規模化については，まだオントロジーに追加する候補を適切に絞れていないという問題点がある．よって，選択のための新たな指標をより多く取りいれることで，その後の上位クラス同定，その他の関係の同定の精度を高めることが求められる．また，Linked Open Data を用いた大規模化については，上位オントロジーの概念と WikiLink で関係づけられている単語のうち，オントロジーに追加すべきものを選別することや，適切な上位クラスの選択などが課題としてあげられる．さらに，現段階では DBpedia Japanese のみの考察に留まっているが，日本語 Wikipedia オントロジー[玉川 11][1]や LODAC Species[2]，Web NDL Authorities[3]といった他の LOD を用いることでより多くの概念や関係を追加することを目指す．

なお，本研究におけるオントロジーの大規模化は，「昆虫」の多様性を模倣したバイオミメティクス研究の知識を対象として行ったが，他の生物種についても同様の手法を順次適用していく．

謝辞

本研究の一部は科学研究費補助金　新学術領域（研究領域提案型）24120002「バイオミメティクス・データベース構築」の助成による．

参考文献

[北河 13]北河 祐作: 大規模化オントロジーの知的探索に向けた多段階展開型概念検索システムの開発，人工知能学会研究会資料, SIG-SWO-A1203-09, 2013.

[古崎 13] 古崎 晃司: 生物多様性を規範とした材料技術開発支援に向けたバイオミメティック・オントロジーの試作，2013 年度人工知能学会全国大会, 3I1-3, 2013.

[古崎 14] 古崎 晃司: オントロジーと Linked Data に基づくバイオミメティック・データベースの構築，2014 年度人工知能学会全国大会, 2F1-4, 2014

[下澤 08] 下澤 楯夫: 昆虫ミメティックス~昆虫の設計に学ぶ~, NTS, 2008.

[下村 10] 下村 政嗣: 生物の多様性に学ぶ新時代 バイオミメティック材料技術の新潮流，科学技術動向 Vol.110, pp.9-28, 2010

[多田 14] 多田 恭平: 専門文書と Linked Open Data を用いたバイオミメティクス・オントロジーの大規模化の試み，2014 年度人工知能学会全国大会, 2F1-5, 2014

[玉川 11] 玉川 奨: 日本語 Wikipedia からプロパティを備えたオントロジーの構築，人工知能学会論文誌，Vol26, No.4, pp504-517, 2011.

[鳥村 15] 鳥村 匠: 生物の機能実現方法に基づく発想支援のためのオントロジー構築とそのガイドラインの提案，2015 年度人工知能学会全国大会, 2M1-5, 2015

[ヒース 13] トム・ヒース: Linked Data: Web をグローバルなデータ空間にする仕組み, 近代科学社, 2013

[1] http://www.wikipediaontology.org/

[2] http://lod.ac/

[3] http://id.ndl.go.jp/auth/ndla/