2F1-5

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

A Trail of Expansion of a Biomimetic Ontology using Technical Documents and Linked Open Data

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

[*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 very important to develop biomimetic database which can bridge various knowledge among through different domains for expansion of biomimetic researches. Biomimetic ontologies makes up the core of the database because it provide systematized knowledge through domains. However, targets of the biomimetic ontology is too large to develop it by hands. This article proposes methods to expand a biomimetic ontology from two approaches. The one is based on natural language techniques for technical documents and the other uses Linked Open Data published on the Web.

1. はじめに

異分野融合を目指した次世代科学の実現には，各分野における多種多様なデータや知見の，分野を超えたスムーズな連携が不可欠である．その中で，生物がもつ優れた機能や製造プロセスを模倣し，技術開発やものづくりに生かそうとする分野，「バイオミメティクス」への注目が高まっている．バイオミメティクス研究を支えるオープンイノベーションプラットフォームを構築するためには，種々の異分野データと知識の相互運用性の確保が必須となる．バイオミメティクス分野の大規模なオントロジー構築は，新たな技術を開発しようとしている工学研究者が，昆虫，鳥類，魚類などの生物多様性と適応に関する情報をあらゆる角度から検索することを可能にし，技術革新の着想を得ることができる「発想支援型のデータベースシステム」を構築することに貢献する．そして，多種多様な生物の情報を幅広く扱うことが必要となるバイオミメティクス研究の特徴から，バイオミメティクス・オントロジーを大規模化することが必要となっている．そこで本研究では，先行研究[古崎 13]を基に，オントロジーの大規模化の手法を中心に研究を行った．大規模化の手法としては，専門文書を対象とした自然言語処理技術の利用，および，再利用可能な形で Web 上に公開されている構造化されたデータベースである Linked Open Data の利用[ヒース 13]という 2 つのアプローチから検討した．なお，本研究におけるオントロジー拡充は，「昆虫」を模倣したバイオミメティック研究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適用することができる．

以下，2 章ではバイオミメティクス・オントロジーを本研究方針で構築するにあたって必要となるバイオミメティクス・上位オントロジーの概要とその構築結果について述べる．3 章では，バイオミメティクス・上位オントロジーを利用したバイオミメティクス・オントロジーの拡充について，その概要と構築結果を述べる．4 章では，本研究の総括と今後の展望について述べる．

2. バイオミメティクス・上位オントロジー

2.1 バイオミメティクス・上位オントロジーの概要

バイオミメティクス・オントロジーを用いたバイオミメティクス・データベースによって幅広い生物種の多様性を適切に扱うためには，数多くの概念をもったオントロジーにする必要がある．生物種の数については主要なものだけでも数万，詳細なものも含めると数百万に至る．それだけ多くの概念を手動でオントロジーに追加するのは不可能に近い．そのため，概念の追加を自動的に行う必要がある．しかしながら，オントロジーを構築するにあたり，数多くの概念を自動的に追加するためには，拡充の指標となる最低限の上位オントロジーを人間の手であらかじめ用意しなければならない．本研究では，その上位オントロジーをバイオミメティクス・上位オントロジーと呼ぶ．

バイオミメティクス・上位オントロジーの構築には，昆虫の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]を用いた．この本には，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，そのキーワードを集めて索引化した「キーワード索引」が記載されている．本研究では，そのキーワード索引に掲載されている語をバイオミメティクス・オントロジーの上位オントロジーに追加されるべき概念として，手動でオントロジーを構築した．

2.2 上位概念の決定

キーワード索引には，カメムシやハエトリグモといった虫の名前，飛行や捕食といった虫の行動，頑強性や走行性といった虫の性質，アドレナリンやフェロモンといった生物が発する物質のような様々な語が全部で 657 語記載されている．バイオミメティクス・データベースで用いるためには，単語同士の関連を示すようなメタデータを付与しなければならない．そこで，キーワード索引に含まれる 657 語に対して，その上位となる概念を付与した．以下に付与した上位概念と，その下位概念となる語の数を示す．

上位概念……下位概念となる語の数
- ・器官……106 語
- ・機能……115 語
- ・機能物……50 語
- ・現象……26 語
- ・構造……12 語
- ・自然物……3 語
- ・時代……2 語
- ・人工物……25 語
- ・性質……63 語
- ・生体行動……57 語

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

・動物の種類……9 語
・動物の名前……25 語
・物質……5 語
・物質(自然)……45 語
・方法……6 語
・理論……18 語
・その他……80 語

これを見ると，機能や機能物，生体行動，また器官，性質を上位概念に持つ概念が比較的多く現れていることが分かる．これにより，先行研究で述べられているように，

・機能 → 生物種 → 構造
・機能 → 生体環境 → 生物種 → 構造
・機能 → 生物の行動 → 生物種 → 構造
・機能 → 構造 → 生物種

といったような，機能を中心とした様々な観点からのつながりが見え，これらの概念のつながりを利用し，「ある機能を実現している「生物(の部位)」の検索」などを可能とするために必要なオントロジーを構築する際に，必要となる概念を多く含んでいると思われる．

3. バイオミメティクス・オントロジーの拡充

3.1 オントロジー拡充に向けた 2 種類のアプローチ

生物規範工学の専門的知識をオントロジーに拡充するために，自然言語処理と Linked Open Data を用いた 2 つのアプローチを提案する．前者においては，前章でも用いた「昆虫ミメティックス」をオントロジー拡充においても用いることを考える．本書は 1000 ページ近くある書籍であり，昆虫以外の領域にも適用な拡充手法を目指すことを考えると，オントロジー構築者が全文書を読んでオントロジー拡充に必要な知識を人手で抽出することは現実的ではない．そこで本研究では，「昆虫ミメティックス」の本文に対して自然言語処理技術を適用することで，オントロジーに必要な概念を追加するという方針を採用した．

また後者においては，Web 上で，計算機による知識処理に適した RDF フォーマットを用いて構造化したデータを，誰もが利用できるオープンデータとして公開されている Linked Open Data を用いることを考える．本研究では，一般的な百科事典として Web 上で構築されている Wikipedia の情報を元に構築された DBpedia Japanese[1]および日本語 Wikipedia オントロジー[玉川 11][2]を用いることとした．

これらの両アプローチに共通する手順は，下記の 3 ステップとなる．

(1)オントロジーに追加する概念の候補を選択する．
(2)オントロジーに追加する概念の上位概念を同定する．
(3)その他の関係の種類を同定する．

次項以降はこの 3 ステップに沿って説明する．

3.2 自然言語処理を用いたオントロジー拡充

自然言語処理を用いたオントロジーの拡充の手順は以下の通りである．

(1)オントロジーに追加する概念の候補の選択

拡充するオントロジーに追加する概念候補を選択するにあたり，まず，テキストマイニング用ソフトウェア「Text Mining Studio」[3]のテキスト文中に出現する単語を頻度順に出力する機能を用いて，「昆虫ミメティックス」本文中にある単語のうち高頻度で出現しているものを抽出した．さらに重要な単語を絞り込む際には，「昆虫ミメティックス」の巻末に掲載されている索引の語を追加する概念の候補とする．索引語は書籍の著者である専門家によって選定されているため，専門家の観点から重要と判断された語であるとみなすことが出来る．

(2)オントロジーに追加する概念の上位概念の同定

前章で述べた上位オントロジーで定義されている概念には，上位概念が既に定義されている．同じ種類の概念は上位概念が同じとなるので，上位オントロジーで既に定義されている概念と類似度が高い単語は，上位概念が同じ兄弟概念になると考えられる．そこで(1)で選択したオントロジーに追加する候補となる単語と上位オントロジーの概念の間の単語間類似度を計算し，この類似度が高いものをオントロジーに追加する．例えば，上位概念が器官として定義されている上位オントロジーの概念「平衡胞」と，それと類似度が高い単語「尾扇肢」の組がある場合，「尾扇肢」の上位概念も器官としてオントロジーに追加する．

単語間の類似度を計る尺度は様々あるが，本研究では，自然言語処理で一般的によくつかわれているコサイン類似度を用いた．コサイン類似度とは，特徴ベクトル間の距離に基づく尺度[相澤 07]であり，各単語をそれぞれベクトル化して表すことによって，比較する 2 つの単語を表すベクトルが類似していればその単語同士も類似しているという考え方のもとで計算される．本研究では，類似度計算のために必要な特徴ベクトルとして，(1)で抽出した「昆虫ミメティックス」の全本文中に高頻度で出現した単語について，1 文毎の共起語を抽出し，書籍の全本文を通した和集合を特徴ベクトルとした．計算を簡単にするため，単語の共起回数は考慮せず，ある単語が書籍全体を通して 1 回でも同一文章に共起して現れていれば特徴ベクトルにおけるその共起語に対する値は“1”，そうでなければ“0”とした．

(3)その他の関係の同定

その他の概念間の関係は，書籍の本文から抽出された単語間の「共起関係」および「係受け関係」を用いる．ただし，これらの関係が抽出された単語間には，何らかの関係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，関係の種類については，共起と係り受けから得られた関連性のある単語のそれぞれの上位概念を比較することで同定する．

例えば，単語 A(例:多層膜干渉)と単語 B(例:ナノ構造)に関連性が認められ，それぞれの上位概念が「機能」と「構造」であった場合，それらの関係は「機能達成に貢献する構造」と考えられる(図 1)．

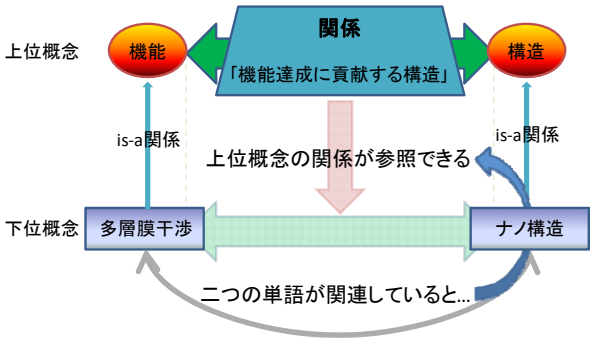

図 1 その他の関係の同定の例

[1] http://ja.dbpedia.org/
[2] http://www.wikipediaontology.org/
[3] https://www.msi.co.jp/tmstudio/

表 1「昆虫ミメティックス」に出てくる全ての単語のコサイン類似度による上位概念同定の正答率

コサイン類似度	全組	判定数	正しい組	誤っている組	正答率
0.95-1.00	10	10	3	7	30.0%
0.90-0.95	13	13	6	7	46.2%
0.85-0.90	23	16	3	13	18.8%
0.80-0.85	20	10	4	6	40.0%
0.75-0.80	33	12	5	7	41.7%
0.70-0.75	55	11	2	9	18.2%
0.65-0.70	146	14	4	10	28.6%
0.60-0.65	450	12	2	10	16.7%
0.55-0.60	1628	22	3	19	13.6%
0.50-0.55	4942	24	3	21	12.5%

表 2　事項索引の単語のコサイン類似度による上位概念同定の正答率

コサイン類似度	全組	判定数	正しい組	誤っている組	正答率
0.80-1.00	8	8	2	6	25.0%
0.70-0.80	11	11	7	4	63.6%
0.60-0.70	53	11	6	5	54.5%
0.50-0.60	371	20	6	14	30.0%

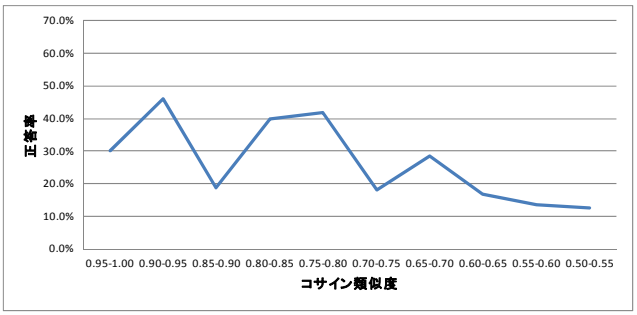

図 2　抽出した全単語を対象としたコサイン類似度による上位概念同定結果の評価

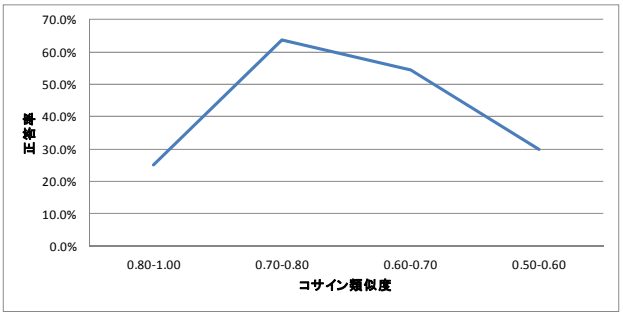

図 3　事項索引の単語を対象としたコサイン類似度による上位概念同定結果の評価

3.3 自然言語処理を用いたオントロジー拡充の考察

まず,「昆虫ミメティクス」の全本文 16,743 文を「Text Mining Studio」を用いて処理し, 出現回数が 2 回以上の単語 10,616 語を抽出した. 次に, この 10,616 語の単語それぞれについて, 同じ文中に共起して出現する単語の組を導出した. 得られた共起する単語の組は 2,134,511 組であった(異なる文に現れた同じ単語の組み合わせも含む). さらに, その結果を用いて単語毎に共起して出現する語の一覧を生成した. これがその単語の特徴ベクトルに相当する. そして, その共起語の組から生成した各単語の特徴ベクトルを用いて, 各単語間のコサイン類似度を計算した. あまり類似度が高くない単語の組の上位概念を同じにするのは不適切であるため, 本研究では, コサイン類似度が 0.5 以上の単語の組 8731 組について考察を行った.

まず最初に, コサイン類似度が 0.5 以上の単語の組 8731 組のうち, 追加候補の単語がキーワード索引と重複していないもの 7320 組について, コサイン類似度が高い順にソートし, 単語の組をランダムに選び, まだ上位概念が付与されていない事項索引の上位概念にその単語と類似している上位オントロジーの概念の上位概念を付与したときの正答率とコサイン類似度の関係を手動で調べた. その結果を表 1 に, グラフを図 2 に示す. これによると, 正答率が高くても 46.2%ほどしか得られなかった. これは, 全ての高頻度で出現する単語があまり重要でない単語やうまく抽出出来ていない単語が含まれているためだと考えられる. そこで,「昆虫ミメティックス」に付録されている, 専門語の索引である事項索引の単語に限定して同様の処理を試みた.

事項索引には 2228 語の単語が掲載されているが, この中にはキーワード索引と重複した単語も含まれている. 上位オントロジーの概念と類似した事項索引の単語の組のうち, 事項索引がキーワード索引と重複していない単語の組でコサイン類似度が 0.5 以上であるものは 443 組であった. これに同様の処理を行い上位概念を付与したときの正答率とコサイン類似度の関係を調べた. その結果を表 2 に, グラフを図 3 に示す. これによると, コサイン類似度によって正答率が 63.6%にまで高めることが出来た. このことから, 追加候補とする単語を適切に限定することが正答率をあげることに重要であることが示唆される.

また, コサイン類似度が高くなりすぎると, グリア細胞(上位オントロジーの概念)とグリア(事項索引の単語)や捕食(上位オントロジーの概念)と捕食者(事項索引の単語)のように, どちらか一方がもう一方の単語の一部分になってしまっている場合が多くなり上位概念が違う場合があったが, 全体的に見るとコサイン類似度が低くなるに従って上位概念が同じになる確率が低くなっていっている.

3.4 Linked Open Data を用いたオントロジーの拡充

Linked Open Data を用いたオントロジーの拡充の手順は以下の通りである.

(1)オントロジーに追加する概念の候補の選択

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」[1]を用いて，前章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のうち DBpedia Japanese と日本語 Wikipedia オントロジーに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係をもつデータをオントロジーに追加する概念の候補とした．

(2)オントロジーに追加する概念の上位概念の同定

DBpedia Japanese，日本語 Wikipedia オントロジーのデータと一致する概念に関しては，それぞれの LOD に定義されている関係を利用できる．それらの関係を用いて導入された追加する概念の候補の上位概念については，上位オントロジーで定義されているトップレベルの上位概念に相当するデータが既に各 LOD で定義されている場合は，それらとの関係を調べることで同定できる．それ以外の場合は，各 LOD 内の上位概念相当のデータと上位オントロジーのマッピングを行うなどの工夫が必要である．

(3)その他の関係の同定

その他の関係の同定については，自然言語処理による手法と同様に，上位概念の同定結果を用いる手法に加え，各 LOD で定義されている関係の種類を利用する方法が考えられる．

3.5 Linked Open Data を用いたオントロジー拡充の考察

「簡易 SPARQL ツール」によって，上位オントロジーの概念 657 個のうち，DBpedia Japanese にデータが存在するものは 285 個（43.4%），日本語 Wikipedia オントロジーにデータが存在するものは 226 個（34.4%）であることが分かった．これらの概念に関しては，日本語 Wikipedia オントロジーを用いた is-a 関係やその他の関係にある概念の追加や，DBpedia Japanese を用いた上位オントロジーの概念と関連性のある単語の追加が行える．

また，前章で定義した上位オントロジーの上位概念についても同様の解析を行った．その結果，その他を除いた上位概念全 16 個のうち，器官，機能，現象，構造，時代，性質，物質，方法，理論の 9 個の概念に DBpedeia Japanese と日本語 Wikipedia オントロジーのデータが存在することが分かった．このデータを用いることにより，今後追加する概念が既に DBpedia Japanese や日本語 Wikipedia オントロジーに上位概念の単語と is-a 関係やその他の関係にあると定義されている場合にそのままその関係を用いることが出来る．しかし，器官や物質といった上位概念は，本来の意味が生物や実際に存在しているものに関したものであるため，LOD によって定義された関係を用いることが可能であろうが，機能や構造に関しては，構築したいオントロジーがバイオミメティクスに関係するものであるため，追加したい概念が関係づけられていない可能性がある．これらの LOD に存在しない上位概念については，オントロジーのマッピングを行うなどして対応を検討したい．

一方，DBpedia Japanese の WikiLink を用いた関連概念の追加に関しては，特性上，概念間の関係を同定する必要があるため，Wikipedia 記事内での自然言語処理での類似度計算による関係の導出などの操作を行わなければならない．これについては今後の研究によって効率的な関係同定の方法を考える必要がある．

4. 本研究の総括と今後の展望

本研究では，自然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，生物学と工学の両分野の研究者がそれぞれの多種多様な知識や知見，データを領域横断的に利用できるようにするための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の手法についての考察を行った．

今後の研究では，様々な観点からの検索を可能としたデータベースの構築のためにはより多くの知識をオントロジーに拡充する必要があり，これを多角的に扱うためには，オントロジーの概念数を多くすることに加えて，概念間により多くの関係を持たせることが重要となってくるため，今回提案した 2 つの概念追加のアプローチをより考察することで，生物学と工学という全く異なる分野の連携を促進していきたい．また，自然言語処理を用いたオントロジーの拡充については，コサイン類似度による上位概念同定の正確性がまだ不足しているため，今回行った 1 文毎の特徴ベクトルの計算を段落毎，同一ページごとなどに変更し，どの範囲で特徴ベクトルを生成すれば良いかの検討をする必要がある．さらに，Web 検索エンジンを用いた類似度計算[Bollegala 07]など，コサイン類似度以外の手法を取り入れることも考察したい．また，Linked Open Data を用いたオントロジーの拡充については，DBpedia Japanese，日本語 Wikipedia 以外にも日本語 WordNet[2]などの既存のシソーラスを用いること[Morita 08，山口 99]も検討している．

謝辞

本研究の一部は科学研究費補助金 新学術領域研究（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

[相澤 07] 相澤彰子:共起に基づく類似性尺度，オペレーションズ・リサーチ 経営の科学，Vol.52，No.11，pp.706-712，2007.

[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.

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

[玉川 11] 玉川 奨，森田 武史，山口 高平: 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，Vol.26，No.4，pp.504-517，2011.

[ヒース 13] トム ヒース (著)，クリスチャン バイツァー (著)，武田英明 (監訳)：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

[Bollegala 07] D. Bollegala, Y Matsuo and M Ishizuka: Measuring Semantic Similarity between Words Using Web Search Engines, Proc. of the 16th international conference on World Wide Web, pp. 757-766, 2007.

[Morita 08] T. Morita, N. Izumi, T. Yamaguchi: Integrating a Domain Ontology Development Environment and an Ontology Search Engine, Proc. of the Eighth Joint Conference on Knowledge-Based Software Engineering, Frontiers in Artificial Intelligence and Applications pp.263–272, 2008.

[山口 99] 山口高平，樽松理樹，青木千鶴，関内律恵子，加賀山茂，吉野一：計算機可読型辞書を利用した領域オントロジー構築支援環境，人工知能学会誌，Vo.14,No.6，pp.1080-1087，1999.

[1] http;//sourceforge.jp/projects/easylod/wiki/EasySPARQL

[2] http://nlpwww.nict.go.jp/wn-ja/