5B-06

Wikipediaから構築した生物オントロジーによる映像コンテンツの体系的な表現方法の検討

浦川 真† 宮崎 勝† 山田 一郎† 中川 俊夫† 藤沢 寛†

NHK放送技術研究所†

1 はじめに

インターネットを経由した映像配信サービスの普及に伴い、様々な分野での映像コンテンツの利活用が期待されている。利活用において、映像の内容を示すメタデータが重要となる。メタデータを介した映像の利活用として、放送番組と字幕情報を用いたオンライン百科事典生成[1]や、ニュース映像と日付情報を用いたWikipediaの拡張[2]などが提案されている。

本稿では、Wikipediaの見出しや本文から、生物を説明する際に必要となる構造を生物オントロジーとして構築し、このオントロジーを用いた新しい映像コンテンツ利活用手法を提案する。構築した知識構造により、映像を生物の体系的な説明に利用できるだけでなく、Wikipediaといった外部サービスとの補完連携など幅広い展開も可能となる。

2 Wikipediaからの生物オントロジーの構築

インターネット上の集合知であるWikipedia記事は、見出し構造により説明が体系化されている一方で、その書き方は統一されていない。例えば、ミジンコは、「特徴」「生息環境」「生態」といった見出しで説明されているが、クモは、「体の構造」「雄雌」「内部形態」「生活史」といった見出しで説明されている。そこで、生物に関する複数の記事から、見出しとその特徴を抽出することで、生物の説明に共通利用できる知識構造を構築する。

2．1 知識構造の抽出

動物界に属する161の生物について、見出しとその本文から、以下の処理により生物オントロジーを構築した。①多くの生物記事で利用されている見出しを抽出する。②各見出しを、本文内に出現する、ガ格、ハ格、ノ格の名詞と動詞によりベクトルで表現する。ベクトルの要素の値は単語の出現頻度とする。③2つの見出しベクトルのコサイン類似度が0.7以上の場合は統合する。④見出しの上位下位関係を参考に、手作業により階層化する。

Wikipediaから抽出した見出しの階層構造を図1に示す。「生活環」は最上位見出しであるが、「生態」の下位見出しである「生活史」との類似度が0.7以上のため統合した。なお、Wikipediaでの各生物記事内の平均見出し数は2.3のため、第3階層の「外部形態」「内部形態・内部構造」「分布」「食性」「生活史・生活環」の5つの見出しを、概念タグとして生物を説明するために利用する。概念タグで使用される単語の中からスコア値が高い単語を抜粋して表1に示す。生物を説明する際に必要となる概念タグや、概念タグを特徴付ける単語を取得することができた。

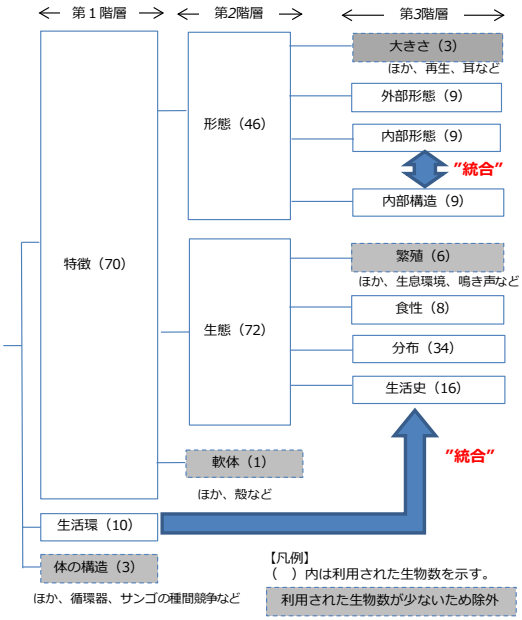


図1 抽出した概念関係

表1 概念タグと特徴的な単語

概念タグ	各概念タグを特徴づける名詞
外部形態	腕、足、体、繊毛
内部形態・内部構造	管、腸、心臓、神経
分布	日本、群、地、周辺
生活史・生活環	卵、幼虫、蛹、交尾
食性	性、植物、獲物、食物
概念タグ	**各概念タグを特徴づける動詞**
外部形態	ある、並ぶ、出る
内部形態・内部構造	構成する、結合する
分布	分布する、生息する
生活史・生活環	なる、孵化する、成長する、羽化する
食性	食べる、捕食する、剥がす、噛みつく

A study on systematical view of video content based on biological ontology from wikipedia

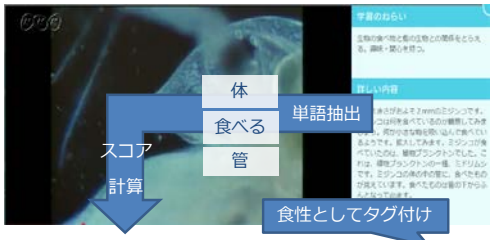
†Makoto URAKAWA †Masaru MIYAZAKI †Ichiro YAMADA †Toshio NAKAGAWA †Hiroshi FUJISAWA

†Science & Technology Research Laboratories, Japan Broadcasting Corporation(NHK)

3　検証

3．1　映像コンテンツへの概念タグ付与

構築した知識構造を利用し、NHK の学校向け WEB コンテンツである NHK for School[1]で公開されている、生物に関する映像コンテンツ 371 本への概念タグ付与実験を行った。映像には、その映像内容がテキストで付与されている。各概念タグを特徴づける単語が、映像内容テキストに出現した場合に、該当する概念タグに加点し、そのスコアが最も高い概念タグをコンテンツに付与した（図 2）。



	内部構造・形	外部形態	分布	生活史・生活	食性
「体」	0.44	0.89	0.00	0.50	0.00
「管」	1.31	0.00	0.00	0.09	0.00
「食べる」	0.00	0.00	0.00	0.35	1.25
⋮					
合計	3.52	3.68	0.32	7.09	8.50

図 2　コンテンツ毎のスコア計算例

手動で付与した各コンテンツの概念タグと、本検証で得られた概念タグを比較した結果を表 2 に示す。Wikipedia から構築した知識構造では、「外部形態」を特徴づけるべき「眼」が、「内部構造・内部形態」のスコアの方が高かったため、適合率の精度を下げる結果となった。

表 2　適合率と再現率

概念タグ	適合率	再現率	F 値
内部構造・内部形態	0.38	0.67	0.48
外部形態	0.91	0.78	0.84
分布	1.00	0.50	0.67
生活史・生活環	0.81	0.89	0.85
食性	0.74	0.70	0.72

3．2　Wikipedia と映像コンテンツの紐づけ

例として、「ミジンコ」の Wikipedia 記事の見出しに対し、2 章で構築した知識構造を用いて新たな概念タグを付与し（表 3）、映像コンテンツとの紐づけを行った。3.1 で概念タグを付与した NHK の映像コンテンツと、Wikipedia 記事とが補完連携するイメージを図 4 に示す。表 3 に示すように、Wikipedia 記事には「内部構造・内部形態」「食性」に関する記述がないことが分かる。図中の赤色のテキストは、映像コンテンツにしかない情報で補完することを表している。このように、映像コンテンツを単純に紐づけるのではなく、体系化された知識に基づき連携させることで、より網羅性のある生物情報の説明を、映像付きで行うことができる。

表 3　Wikipedia 記事（ミジンコ）の再定義

Wikipedia 記事		付与した概念タグ
見出し	本文	
特徴	中型種で体長 1.5-3.5㎜。体は頭部を除き二枚貝のような背甲で覆われ、横から見る・・・	外部形態
生息環境	世界的に分布する。日本でも全土に分布、浅い池沼に生息する。	分布
生態	ミジンコには、自分とおなじクローンしか産まない単為生殖期と、交配して子孫を残す有性生殖期がある。一般的に、・・・	生活史
利用	ミジンコの遺伝子は 3 万 1000 個以上にのぼり、ヒトよりも 8000 個も多い。	該当なし

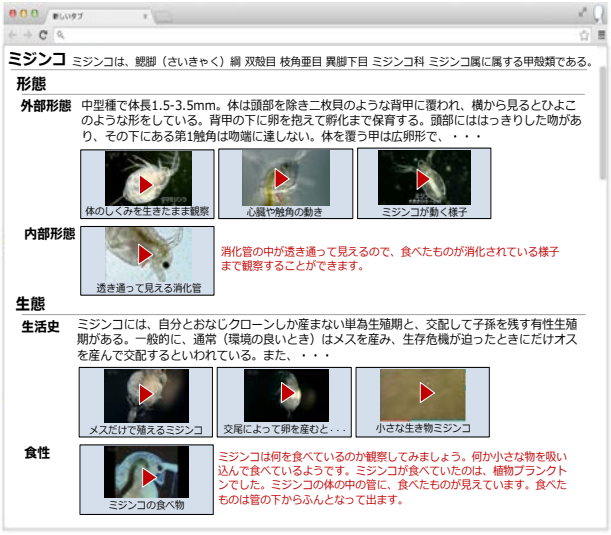


図 4　映像コンテンツと Wikipedia の連携案

4　まとめ

生物を説明するために必要な知識を、生物オントロジーとして構造化することにより、映像コンテンツの体系化や、Wikipedia などの外部サービスとの補完連携が可能となることを示した。今後、「内部構造・内部形態」は“体内の臓器”に関する単語を持つといった知識を利用した推論により精度向上に取り組み、より高度な知識構造化を検討していく。さらに、映像コンテンツの利活用に向けたオープンデータ化にも取り組んでいく。

[1] http://www.nhk.or.jp/school/

参考文献

[1] Miura et al.,“Automatic Generation of a Multimedia Encyclopedia from TV Programs by Using Closed Captions and Detecting Principal Video Objects”, Proc . 8th IEEE ISM , pp. 873 － 880 , Dec . 2006

[2] 奥岡ほか：ニュース映像間の意味構造を利用した Wikipedia 情報の拡張，情報処理学会第 72 回全国大会講演論文集， pp.5-195-5-196(2010).