

# Wikipedia を用いた多言語情報アクセスに関する研究: 言語間リンクの分析と応用

## Crosslingual Information Access using Wikipedia: Analysis and Application of Interlanguage-links of Wikipedias

新井 嘉章<sup>1</sup> 福原 知宏<sup>2</sup> 増田 英孝<sup>1</sup> 中川 裕志<sup>3</sup>

<sup>1</sup> 東京電機大学 未来科学部

<sup>1</sup>School of Science and Techonology for Future Life, Tokyo Denki University

<sup>2</sup> 東京大学 人工物工学研究センター

<sup>2</sup>Reseach into Artifacts, Center for Engineering, The University of Tokyo

<sup>3</sup> 東京大学 情報基盤センター

<sup>3</sup>Information Technology Center, The University of Tokyo

**Abstract:** Interlanguage-links (ILLs) among Wikipedias are one of important multilingual resources. In this paper, we describe (1) an analysis results of ILLs among Chinese, Japanese, Korean, and English (CJKE) Wikipedias, (2) evaluation results of ILLs using traditional dictionaries, (3) hierarchic analysis about Category-links of Wikipedias, (4) our cross-lingual keyword navigation system using ILLs.

## 1 はじめに

ウェブ上のコンテンツは、母国語が異なる様々な利用者によって作成されている。その為、ウェブ上のコンテンツで使用されている言語は多様化しており、Wikipedia によれば、ウェブ上のコンテンツは、英語 56.4%、ドイツ語 7.7%、フランス語 5.6%、日本語 4.9%で、75以上の言語が使用されていると言われている<sup>1</sup>。しかし、現在、検索エンジンなどを用いたウェブへのアクセスでは、利用者の母国語や習得言語などを用いた、キーワード入力しか行われず、全てのコンテンツにアクセス出来ていない現状がある。

本研究では、このような多言語情報へのアクセス支援という課題に対して、Wikipedia の言語間リンクを利用した多言語対訳システムを構築し、さらに、得られる訳語数や、訳語の質を調査する為に、Wikipedia の言語間リンクの詳細な分析を行った。また、本研究では、多言語情報へのアクセス支援という課題に対し、Wikipedia を多言語情報資源として用いる為、Wikipedia の各言語版がどのように異なっているのか知る必要がある。その手がかりになる調査として、本研究では、Wikipedia

のカテゴリに関する分析を行い、カテゴリ階層の深さなどの、言語毎での違いを明らかにした。

本論文の構成は次の通りである。2 では、Wikipedia の関連研究を紹介する。3 では、Wikipedia の言語間リンクの分析について述べる。4 では、Wikipedia のカテゴリリンクの分析について述べる。5 では、言語間リンクを用いた多言語対訳システムについて述べる。6 では、考察を述べる。7 では、まとめと今後の課題について述べる。

## 2 Wikipedia に関する研究

Wikipedia に関する研究には主に、情報利用・教育に関する研究、自然言語処理に関する研究、Wikipedia のコミュニティを分析した研究がある。以下に、いくつかの研究と、本研究との関連を示す。

### 2.1 情報利用・教育に関する研究

Wikipedia をオントロジ構築に活用する研究に、中山による ‘Wikipedia マイニングによる大規模 Web オントロジの実現’[1]、Syed らの ‘Wikipedia as an Ontology

<sup>1</sup>[http://en.wikipedia.org/wiki/World\\_Wide\\_Web#Statistics](http://en.wikipedia.org/wiki/World_Wide_Web#Statistics)

for Describing Documents'[2]がある。また、Cuiらのオントロジ構築に向けたコーパス作成に Wikipedia を利用した研究 [3] や、Wikipedia を用いたオントロジの評価に関する研究 [4] がある。Milne は、Wikipedia を用いて意味的関連性 (Semantic Relatedness) を計算する為の手法 'Wikipedia Link Vector Model (WLVM)' を提案した [5]。この他に、情報利用に関する研究として、Wikipedia の情報と地図情報を、タッチパネルを用いて連動させる試み [6] がある。教育に関するテーマでは、オンライン教育ツールとしての利用提案 [7, 8, 9] がある。

このように、Wikipedia は知識抽出分野で資源として注目を集めており、オンライン百科事典という本来の利用法とは異なる応用が成されている。一方で、教育現場に Wikipedia という新しい形態のオンラインツールを如何に取り込んでいくかといった、Wikipedia そのものを活用する取り組みもある。

## 2.2 Wikipedia を用いた自然言語処理研究

固有表現辞書の構築や維持には、時間と労力がかかる。その為、従来からブートストラップ手法などを用いた、大量のテキストから自動的に辞書を構築する研究 [10] が行われてきた。Kazama らは、Wikipedia の記事の定義文を用いて、ある単語に関する上位語を返す辞書を構築し、用いることで、英語の固有表現認識の精度が向上することを示した [11]。この他に、Iftene らの、Wikipedia を用いた固有表現の特定に関する研究 [12] や、固有表現の曖昧性解消の為に Wikipedia を用いた研究 [13] がある。

Wikipedia を質問応答システムに利用するタスク [14] が存在する。質問応答システムは、ユーザが日常使っている言葉を用いて、情報検索などを利用できるようにするシステムであり、Sigurbjornsson らは、多言語情報検索の評価用資源として、Wikipedia が有用な事が、今後示されて行くと述べている [15]。

Wikipedia のリンク分析によって、類似記事を特定する研究 [16] がある。また、この研究を多言語に応用し、パラレルコーパスを構築する研究 [17] がある。このように、類似記事や対訳関係の抽出に Wikipedia のリンク情報が用いられている、また、同様にリンク分析から同義語を抽出する研究 [18] もある。

Wikipedia のような大規模かつ無償利用可能なデータ資源の登場は、自然言語処理の分野において、従来手法の精度改善や、大規模な対訳コーパスの獲得など、変化をもたらしている。

## 2.3 Wikipedia コミュニティの分析研究

Viegas らは、コミュニティの発展を理解する上で、編集履歴が重要であるとし、'History Flow' と呼ばれる履歴情報の可視化・分析ツールを開発した [19]。また、統計情報の分析では、Voss の、1 記事あたりの編著者数、1 著者あたりの編集記事数などの分析 [20] や、Almeida らの分析 [21]、Geser による多言語での時系列分析 [22]、Francesco らの、'HITS' と 'PageRank' を用いた分析 [23] がある。Zesch らは、従来のセマンティックネットワークのグラフ理論による分析手法を、Wikipedia のカテゴリグラフ (WCG) に適用し、自然言語処理への応用を提案した [24]。Cifolilli は、仮想コミュニティとして Wikipedia を考察し、落書き投稿対策と、Wiki による編集コスト低減が、Wikipedia がコミュニティとして成功した要因であると述べた [25]。このように、Wikipedia は、Wiki によって発展し成功したが、その一方で、画像リソースに関する作業環境は、提供されておらず、画像の版管理ができないといった指摘 [26] もある。Holloway らは、英語版 Wikipedia の上位 2 階層のカテゴリマップを作成し、Britannica、Microsoft Encarta と比較した [27]。地理に関するカテゴリでは、Wikipedia が 46 に対し、Britannica が 12、Encarta が 13、歴史のカテゴリについては、Wikipedia の 20 に対し、Britannica が 4、Encarta が 10 と、数においては Wikipedia が豊富であり、Wikipedia の優位性が確認されている。Ortega らは、従来の研究に見られた Wikipedia に関する統計情報抽出や、グラフ生成、定量的結果を全ての Wikipedia の言語版で得られるようにする Python スクリプト、'WikiXRay' を開発した [28]。

Wikipedia の研究では、このような分析を通して、コミュニティ活動を理解する研究や、それらの分析を支えるツールの開発が進んでいる。

## 2.4 本研究との関連

Wikipedia に関する研究は、多岐に渡っており、本研究と技術的に関連するものは少ない。対訳のテーマに関しては、パラレルコーパスの構築に関する研究 [17] が近いが、研究としては技術も、目的も異なっている。可視化のテーマに関しては、Viegas らが開発した編集履歴の可視化ツール 'History Flow' [19] や、Andrew による同義語の可視化ツール 'Synarcher' [18] が関連するが、可視化の対象が異なっている。分析に関しては、Wikipedia のリンク構造や、編集履歴を用いたコミュニティ分析 [27] が行われており、Wikipedia のコミュニティ活動を理解しようとする動きが見られる。しかしながら、本研究における言語間リンクに限った詳細な調査や、言語間でのカテゴリ分析などは、行われていない。



図 1: 言語間リンク (日本語版 Wikipedia 'イヌ' のページ)

表 1: 言語間リンクを持つ項目数

|   | 項目数<br>(言語間リンク有りの項目) | 全言語を対象とする<br>言語間リンク数 |
|---|----------------------|----------------------|
|   | 5,836,167            |                      |
| 英 | (895,235 (15%))      | 4,072,516            |
| 日 | (211,390 (26%))      | 2,050,491            |
| 中 | (122,226 (35%))      | 1,536,757            |
| 韓 | (54,797 (58%))       | 1,061,280            |

Wikipedia の記事は、無数のオンラインユーザによって日々メンテナンスされており、オンラインコミュニティにおける多数決による、その時点での合意である。その為、専門家によって、種類や内容が吟味される市販の辞書とは様々な面で異なる。このような背景から、Wikipedia の記事の量や品質を知る為にも、また、本研究のように、Wikipedia を用いて対訳システムを構築する為にも、Wikipedia を分析し、その特性を理解する事は重要となる。

### 3 Wikipedia の言語間リンクの分析

Wikipedia の編著者は、各項目のページ内に、他の言語版 Wikipedia の同一項目へのリンク (言語間リンク) を設定できる。利用者は、言語間リンクを辿ることで、他の言語版の同一項目のページを見ることができる。図 1 に、Wikipedia のページに表示された実際の言語間リンクを示す。本研究では、これらの言語間リンクを Wikipedia のデータから抽出し、対訳情報

表 2: 言語間リンクの各パターン数

|   | パターン              |                  |                     |                  |                  |
|---|-------------------|------------------|---------------------|------------------|------------------|
|   | A                 | B                | C                   | D                | E                |
| 英 | 7,099<br>(2.08%)  | 9,200<br>(2.70%) | 317,971<br>(93.23%) | 331<br>(0.10%)   | 6446<br>(1.89%)  |
| 日 | 14,508<br>(4.79%) | 4,697<br>(1.55%) | 278,281<br>(91.85%) | 271<br>(0.09%)   | 5,208<br>(1.72%) |
| 中 | 16,356<br>(7.88%) | 3,303<br>(1.59%) | 183,958<br>(88.68%) | 1,605<br>(0.77%) | 2,218<br>(1.07%) |
| 韓 | 3,517<br>(2.87%)  | 661<br>(0.54%)   | 114,910<br>(93.84%) | 435<br>(0.36%)   | 2,934<br>(2.40%) |

として用いた多言語対訳システムを構築する。

本節では、(1) 使用した Wikipedia のデータ、(2) 言語間リンクの接続パターン、(3) 既存の対訳辞書を用いた言語間リンクの評価について述べる。

#### 3.1 使用した Wikipedia データ

表 1 に、日本語版 (2007 年 10 月 13 日)、中国語版 (2007 年 10 月 14 日)、韓国語版 (2007 年 10 月 11 日)、英語版 (2007 年 10 月 18 日) から取得した項目数および、全言語を対象とした言語間リンク数を示す。表で示した言語間リンク数から、言語間リンクを用いて、約 100 万から 400 万の訳語が得られる見込みがある事がわかる。

#### 3.2 言語間リンクの接続パターン

我々は、言語間リンクの接続状態を、以下に示す 5 パターンに分類し、92% が相互リンクである事を示した [29]。この結果から、Wikipedia の各言語の項目間は相互に接続し合っており、多言語で対訳抽出可能な状況にある事が確認できた。この為、多言語情報へのアクセス支援に対し、言語間リンクを用いた対訳抽出が利用可能である事がわかる。

以下に、言語間リンクの各パターンの説明を示す。

**Pattern A (単方向リンク)** *Pattern A* は言語 A から言語 B への一方通行の状態である。

**Pattern B (三角リンク)** *Pattern B* は言語 A と言語 B の接続先が一致しない状態である。

**Pattern C (相互リンク)** *Pattern C* は言語 A と言語 B の接続先が一致し、互いに対訳が抽出可能な状態である。

**Pattern D (無効リンク)** *Pattern D* は言語 A から言語 B へのリンクを持つが、言語 B からは言語 A の存在しない項目へリンクしている状態である。

**Pattern E (ミスリンク)** *Pattern E* は言語 A から言語 B の存在しない項目へリンクしている状態である。

### 3.3 既存の対訳辞書を用いた訳語の評価

我々は、言語間リンクを持つ Wikipedia の項目を無作為に 200 件選び、言語間リンクによる訳語を既存の対訳辞書を用いて評価し、辞書と訳語が一致、辞書と訳語が不一致、辞書未登録語に分類した [29]。表 3 に結果を示す、200 件中、日英では 149 件、日中では 170 件、中韓では、189 件が辞書に登録されていないものであり、言語間リンクを用いて、辞書に無い多くの訳語が得られることを示した。この結果より、ウェブ検索のような、新しいキーワードへの対応が必要な分野についても、言語間リンクを用いて、多言語アクセス支援が可能である事がわかる。

表 3: 対訳辞書を用いた訳語の評価

|       | 未登録語        | 訳語一致       | 訳語不一致    |
|-------|-------------|------------|----------|
| 日 ⇔ 英 | 149 (74.5%) | 42 (21.0%) | 9 (4.5%) |
| 日 ⇔ 中 | 170 (85.0%) | 21 (10.5%) | 9 (4.5%) |
| 中 ⇔ 韓 | 189 (94.5%) | 11 (5.5%)  | 0 (0.0%) |

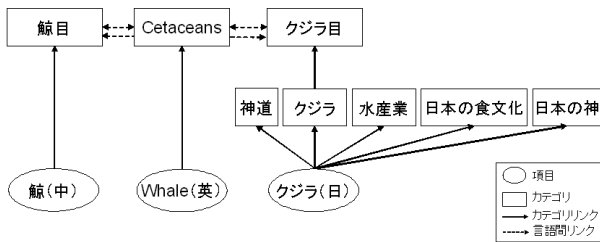


図 2: 言語毎の分類粒度の違い‘クジラ’の例

## 4 カテゴリリンクの分析

Wikipedia の特徴の一つであるカテゴリ (Category) は、MediaWiki の機能の一つであり、自動的に索引を生成する機能である。Wikipedia の編著者は、各記事の本文内に、‘[[Category:動物]]’のようにカテゴリリンクを記述することで、編集対象の記事に対して、Wikipedia 内に存在する任意のカテゴリを設定できる。例として、‘クジラ’には、‘日本の文化’、‘日本の神’、‘神道’、‘水産業’、‘民間信仰’、‘クジラ’のカテゴリが設定されており、様々な観点で分類されている事がわかる。一方、英語版 Wikipedia の ‘Whale’、中国語版 Wikipedia の ‘鯨’では、それぞれ、‘Cetaceans (クジラ目)’、‘鯨目’のみであった。分類の違いは、各言語における観点の違いによっても生じ得るが、図 2 のように、カテゴリ階層の違いによっても起きる。図に示した通り、日本語版の ‘クジラ’ は、英語版、中国語版のように、直接

表 4: カテゴリ数の割合 (日中英韓)

|   | カテゴリ数<br>(カテゴリの割合) | カテゴリを含む全記事数 |
|---|--------------------|-------------|
| 日 | 49,266 (5.1%)      | 957,769     |
| 中 | 38,643 (8.7%)      | 444,152     |
| 英 | 365,210 (5.2%)     | 6,996,745   |
| 韓 | 26,538 (18.9%)     | 140,454     |

‘クジラ目’には属しておらず、英語版、中国語版には存在しないサブカテゴリ ‘クジラ’ に属している。このような、言語間における分類の粒度の違いや、カテゴリの設定状況の違いを明らかにする為に、本研究では基礎的な調査を行った。

本節では、(1) 使用した Wikipedia データ、(2) Wikipedia のカテゴリ階層、(3) 最短経路によるツリー構造への変換、(4) カテゴリ階層の調査、(5) 言語間リンクに関する調査について述べる。

### 4.1 使用した Wikipedia データ

本調査では、日本語版 (2008 年 6 月 7 日)、中国語版 (2008 年 6 月 20 日)、韓国語版 (2008 年 6 月 18 日)、英語版 (2008 年 5 月 24 日) の Wikipedia のダンプデータを用いた。各言語版毎のカテゴリ数と、全記事中における割合を、表 4 に示す、カテゴリ数は、英語版、日本語版、中国語版、韓国語版の順に多く、記事量に比例する傾向が見られる。また、割合は 4 言語中 3 言語が、5% ~ 8% 台だが、記事量の少ない韓国語版のみ 18.9% と高く、記事量に対してカテゴリ数が多い事がわかる。

### 4.2 Wikipedia のカテゴリ階層

Wikipedia では、カテゴリ自体にも、カテゴリを設定できる。例として、‘日本の俳優’には、‘日本の芸能人’、‘各国の俳優’、‘日本の映画’が設定されている。各カテゴリのページでは、それぞれ記事の一覧と、サブカテゴリの一覧を見ることが出来る。このように、Wikipedia のカテゴリリンクは、記事とカテゴリ間だけでなく、カテゴリとカテゴリ間を接続しており、Wikipedia のカテゴリは階層構造を形成している。

日本語版 Wikipedia のカテゴリの上位 2 階層を、表 7 に示す。最上位の ‘主要カテゴリ’ の直下にあるサブカテゴリは、9 カテゴリである。また、英語版 Wikipedia のカテゴリの上位 2 階層を、表 8 に示す。最上位の ‘Contents’ の直下にあるサブカテゴリは、12 カテゴリである。このように、Wikipedia のカテゴリ構成は、Wikipedia の各言語版毎に異なっている。

Wikipedia のカテゴリは、Holloway らによって、ネットワーク分析ツール Pajek<sup>2</sup>を用いた可視化の研究 [27] が行われおり、Zesch らの研究 [24] によって、スケールフリー性、スモールワールド性があること、他の WordNet<sup>3</sup> や Roget's Thesaurus<sup>4</sup>などのセマンティックネットワークと類似している事などが確認されている。また、Zesch らによれば、Wikipedia のカテゴリ構造は、分類学のように厳密に整っておらず、カテゴリ間をループするパターンや、孤立するカテゴリがあるとされているが、稀なケースであるとしている。本調査では、言語間でのカテゴリ階層の比較を容易にする為に、次に述べる最短経路法によって、Wikipedia の複雑なカテゴリネットワークを、ツリー構造に変換する。

### 4.3 最短経路によるツリー構造への変換

Wikipedia のカテゴリは複数の親カテゴリを持つことができる為、各カテゴリについて、トップカテゴリ（日本語版では‘主要カテゴリ’）からの深さを単純に求める事ができない。図 3 にカテゴリリンクの概念図を示す、図の構成では、トップカテゴリの直下に、カテゴリ A、B があり、カテゴリ A の子カテゴリとして、カテゴリ C がある、また、カテゴリ C と B の子カテゴリとしてカテゴリ D がある。この場合、トップカテゴリからカテゴリ D に至る経路は、A-C-D、B-D の 2 経路あり、それぞれ、トップカテゴリからの深さが、3、2 と異なってしまふ。本調査では、各カテゴリについて、図 4 のように、トップカテゴリまでの経路が最短となる親カテゴリを選択する事で、ツリー構造を構築した。

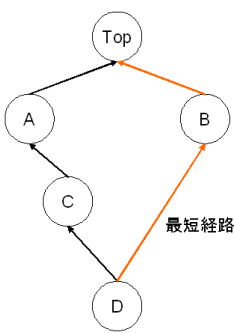


図 3: Wikipedia のカテゴリリンク (矢印の方向はカテゴリリンクの方向を示す)

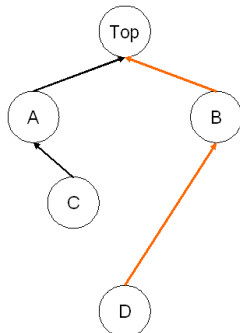


図 4: ツリー構造への変換 (矢印の方向はカテゴリリンクの方向を示す)

表 6: 日本語版 Wikipedia におけるカテゴリの言語間リンク保有率 (階層別)

| 階層 | カテゴリ数 (%)<br>言語間リンクあり | 総カテゴリ数 |
|----|-----------------------|--------|
| 1  | 9 (100%)              | 9      |
| 2  | 203 (78.4%)           | 259    |
| 3  | 1,430 (68.5%)         | 2,089  |
| 4  | 4,600 (67.5%)         | 6,813  |
| 5  | 8,027 (64.7%)         | 12,399 |
| 6  | 5,404 (47.9%)         | 11,275 |
| 7  | 2,130 (36.8%)         | 5,787  |
| 8  | 650 (18.9%)           | 3,448  |
| 9  | 32 (13.9%)            | 230    |
| 10 | 43 (86.0%)            | 50     |

### 4.4 カテゴリ階層の調査

各言語、各階層毎にカテゴリ数を集計した結果を、表 5 に示す。日本語版ではカテゴリの深さは最大で 10 階層、中国語版では 14 階層、英語版では 17 階層、韓国語版では 14 階層となり、また、日本語版、中国語版では 5 階層目、英語版、韓国語版では 6 階層目に位置するカテゴリの数が最も多い事がわかった。

このように、Wikipedia のカテゴリ階層は、言語毎に深さは異なるが、5 階層目、6 階層目に位置するカテゴリが、言語に依らず最も多い事がわかった。次に、各階層毎にどのようなカテゴリが存在するのか調査を行った。表 9 に、カテゴリ‘人間’についてのサブカテゴリの例を階層別に示す。今回の調査では、カテゴリ‘人間’については、上位 7 階層までは種類が豊富だが、8 階層目以降は、ほぼ (9 階層目は全て) 野球選手名に関するカテゴリであり、10 階層目に位置するものは存在しなかった。

尚、日本語版における最も深い階層である、10 階層目に位置するカテゴリは、50 件中 45 件が、‘ヒバリ科’、‘エナガ科’、‘メジロ科’などの鳥に関するカテゴリであり、他 5 件は、‘佐川急便大阪 SC の選手’、‘佐川急便東京 SC の選手’、‘SC 鳥取の選手’、‘バンディオンセ神戸の選手’、‘NBA ファイナル’など、サッカー、バスケットボールなどスポーツ関連のカテゴリであった。

### 4.5 言語間リンクに関する調査

日本語版 Wikipedia のカテゴリについて、言語間リンクの保有率を各階層別に調査した (表 6)。表で示した結果より、上位階層から下位階層にかけて、言語間リンクの保有率が減少している事がわかる (10 階層目を除く) この結果より、下位の階層になるにつれて、言語間リンクを用いた対訳抽出率が低下していく事が明らかとなった。

<sup>2</sup><http://vlado.fmf.uni-lj.si/pub/networks/pajek/>

<sup>3</sup><http://wordnet.princeton.edu/>

<sup>4</sup>[http://en.wikipedia.org/wiki/Roget's\\_Thesaurus](http://en.wikipedia.org/wiki/Roget's_Thesaurus)

表 5: カテゴリ階層 (日中英韓)

| 階層 | カテゴリ数<br>(日)   | カテゴリ数<br>(中)  | カテゴリ数<br>(英)   | カテゴリ数<br>(韓)  |
|----|----------------|---------------|----------------|---------------|
| 0  | 1 (0.0%)       | 1 (0.0%)      | 1 (0.0%)       | 1 (0.0%)      |
| 1  | 9 (0.0%)       | 15 (0.1%)     | 12 (0.0%)      | 3 (0.0%)      |
| 2  | 259 (0.6%)     | 365 (1.6%)    | 162 (0.1%)     | 26 (0.2%)     |
| 3  | 2,089 (4.9%)   | 1,913 (8.3%)  | 2,331 (0.9%)   | 102 (0.9%)    |
| 4  | 6,813 (16.1%)  | 4,488 (19.4%) | 8,929 (3.3%)   | 367 (3.4%)    |
| 5  | 12,399 (29.3%) | 8,258 (35.7%) | 31,418 (11.6%) | 991 (9.2%)    |
| 6  | 11,275 (26.6%) | 4,607 (19.9%) | 67,403 (25.0%) | 6,290 (58.3%) |
| 7  | 5,787 (13.7%)  | 2,346 (10.1%) | 75,246 (27.9%) | 890 (8.2%)    |
| 8  | 3,448 (8.1%)   | 976 (4.2%)    | 51,039 (18.9%) | 950 (8.8%)    |
| 9  | 230 (0.5%)     | 117 (0.5%)    | 23,368 (8.6%)  | 676 (6.3%)    |
| 10 | 50 (0.1%)      | 27(0.1%)      | 6,206 (2.3%)   | 351 (3.3%)    |
| 11 | -              | 11 (0.0%)     | 1,179 (0.4%)   | 152 (1.4%)    |
| 12 | -              | 3 (0.0%)      | 1,002 (0.4%)   | 10 (0.1%)     |
| 13 | -              | 4 (0.0%)      | 1,111 (0.4%)   | 2 (0.0%)      |
| 14 | -              | 4 (0.0%)      | 572 (0.2%)     | 2 (0.0%)      |
| 15 | -              | -             | 327 (0.1%)     | -             |
| 16 | -              | -             | 21 (0.0%)      | -             |
| 17 | -              | -             | 27 (0.0%)      | -             |

## 5 キーワードの多言語対訳システム

### 5.1 言語間リンクの可視化システム

Wikipedia の言語間リンクを可視化するシステムを構築した<sup>5</sup>。図 5 にシステムの画面を示す。本システムでは、利用者が入力したキーワードに関する関連語や、カテゴリ、他の言語への対訳を得られる他、言語間リンクによる、言語間の接続状態を視覚的に確認できる。

図の例では、‘ワイン’を入力語とし、‘Wine’などの訳語が得られた他、‘赤ワイン’、‘白ワイン’、‘葡萄酒’など、同義語を含む関連語を得ている。また、‘ワイン’は、日本語版では、‘調味料’、‘果実酒’のカテゴリに属しているが、中国語版では、日本語版で‘酒’、‘発酵食品’、‘ワイン’にあたるカテゴリに属している事などもわかる。

### 5.2 言語間リンクの多言語ウェブ検索への応用

前節で述べた、キーワードの多言語対訳と、機械翻訳サービスを連携させる事で、多言語ウェブ検索が可能になる。図 7 に、その模式図を示す。図の例では、日本語の‘イルカ’を入力語とし、韓国語、英語、中国語にそれぞれ対訳している。対訳したキーワードを、検索エンジンへの入力とし、得られた各言語による検索結果を、機械翻訳エンジンで翻訳する。このような連携により、多言語検索の結果を日本語で閲覧できる。また、ウェブ検索エンジンに Google を用いることで、入

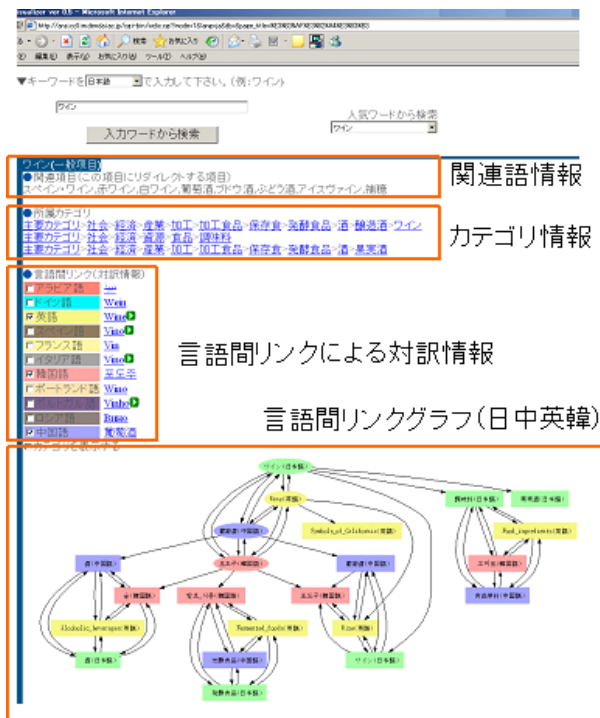


図 5: 言語間リンク可視化システム (http://arai.cdl.im.dendai.ac.jp/)

<sup>5</sup>http://arai.cdl.im.dendai.ac.jp/



図 6: 他サービスとのマッシュアップ例 (<http://arai.cdl.im.dendai.ac.jp/cgi-bin/ltw.cgi>)

力語に関する様々な言語での関連語を見る事ができる。表 10 に実例を示す。日本語版では、防犯製品の‘どこイルカ’や、貸切バス事業者の‘イルカ交通’などがある。韓国語版では、‘イルカのオリンピック’といったゲーム名、英語版では、‘ドルフィン・フィッシュ’といった魚の名前など、言語毎に得られるトピックが異なる事がわかる。

### 5.3 他のウェブサービスとのマッシュアップ

本節では、Google や Flickr のように WebAPI を公開しているサービスを用いた多言語対訳システムの応用を示す。WebAPI を用いると、例えば、Flickr 上で検索した場合に得られる画像の URL や、タグ情報、コメントなどを得る事ができる。本研究で構築した多言語対訳の仕組みを用いれば、多言語でキーワードを指定する事ができ、多言語のコンテンツへのアクセスが可能となる。図 6 に、Google と Flickr の API を用いたシステムを示す。図のシステムでは、画面左上部に Flickr の WebAPI を用いて、得られた画像を表示している。画面左中央部に、Google 検索の結果を表示し、画面右に、Wikipedia のカテゴリ情報、言語間リンクでの対訳を表示している。

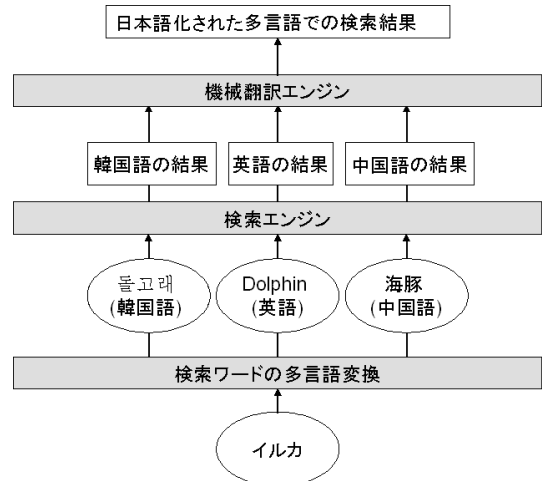


図 7: 多言語検索の模式図‘イルカ’の入力例

## 6 考察

3 では、言語間リンクの接続分類を示し、それらの内訳として 92% が相互リンクである事を述べた。このように、言語間リンクは殆どが相互リンクの状態であり、対訳辞書構築に利用できる。4 では、カテゴリリンクのネットワークをツリー構造に変換する事で、全てのカテゴリについて、トップカテゴリからの階層の深さ一意に求めた。調査では、各言語、各階層毎のカテゴリ数、言語間リンクの保有率を調査した。5 では、言語間リンクを用いて構築した対訳システムを実際に示した。言語間リンクとカテゴリリンクの可視化により、与えられたキーワードに対して、対訳を得られるだけでなく、言語毎でのカテゴリの違いなどを確認できた。それらの違いの分析により、言語間での認識の近さを測る尺度が得られる可能性がある。後半は、検索システムへの応用を示し、機械翻訳サービスと連携した多言語検索システムを示した。Google 検索を用いて得られる関連語は、言語毎に異なり、キーワードに対して、各言語で異なる関連トピックが得られた。このように、一般的なウェブ検索への利用のほかに、このような国ごとの関心の違いを得るような用途に、言語間リンクを用いた、多言語情報検索が利用可能であることを示した。

本研究では、多言語情報アクセスという課題に対して、Wikipedia の言語間リンクは、ウェブ検索システムのような豊富なキーワードに対応できる辞書が要求される分野についても、対訳辞書として用いる事が可能である事を示した。一方で、Wikipedia は各言語によって、各項目に対するカテゴリ情報などが異なり、本研究のような Wikipedia を利用する研究では、Wikipedia の言語毎の特性を、理解しておく必要がある、その為

の調査として、本研究では、カテゴリに関する分析を行った。

## 7 おわりに

本稿では、Wikipediaの言語間リンクの分析結果と、カテゴリリンクに関する調査結果、言語間リンクを活用したシステムについて示した。今後の課題として、Wikipediaの言語資源としての質を高める為に、Wikipediaの各言語版で、分野毎の分類の粒度、項目数、質などを明らかにし、各言語で充実している分野、そうでない分野を明らかにし、コミュニティーにフィードバックしていく事が挙げられる。

## 参考文献

- [1] 中山浩太郎. Wikipedia マイニングによる大規模 web オントロジの実現. 人工知能学会第 22 回全国大会 (JSAI2008), 1I2-01.
- [2] Anupam Joshi Zareen Saba Syed, Tim Finin. Wikipedia as an ontology for describing documents. 2008.
- [3] Gaoying Cui, Qin Lu, Wenjie Li, Yirong Chen. Corpus exploitation from wikipedia for ontology construction. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.
- [4] Jonathan Yu, James A Thom, Audrey Tam. Ontology evaluation using wikipedia categories for browsing. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007.
- [5] David Milne. Computing semantic relatedness using wikipedia link structure. *In Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC 2007)*.
- [6] Schoning, Johannes, Brent Hecht, Martin Raubal, Antonio Kruger, Meri Marsh, Michael Rohs. Improving interaction with virtual globes through spatial thinking: Helping users ask "why?". *Intelligent User Interfaces 2008 (IUI 2008)*, pp.129-138.
- [7] Naomi Augar, Ruth Raitman, Wanlei Zhou. Teaching , learning online with wikis. *Proceedings of the 21st ASCILITE Conference*, pp.95-104, 2004.
- [8] Forte Andrea, Amy Bruckman. From wikipedia to the classroom: exploring online publication and learning. 2006.
- [9] Piotr Konieczny. Wikis and wikipedia as a teaching tool. *International Journal of Instructional Technology and Distance Learning*, 2007.
- [10] Ellen Riloff. Learning dictionaries for information extraction by multi-level bootstrapping. *In Proc of National Conference on Artificial Intelligence*, pp.474-479, 1999.
- [11] Jun 'ichiKazama, Kentaro Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. *In Proc.EMNLP-CoNLL*, pp. 698-707, 2007.
- [12] Adrian Iftene, Alexandra Balahur-Dobrescu. Named entity relation mining using wikipedia. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.
- [13] Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. *In Proc. EMNLP-CoNLL*, pp.708-716, 2007.
- [14] WiQA. 2006. question answering using wikipedia. URL: <http://ilps.science.uva.nl/WiQA/>.
- [15] Borkur Sigurbjornsson, Jaap Kamps, Maarten de Rijke. Focused access to wikipedia. *In Proceedings DIR-2006*.
- [16] Fissaha Adafre, Sisay, de Rijke, Maarten. Discovering missing links in wikipedia. *In Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*.
- [17] Sisay Fissaha Adafre, Maarten de Rijke. Finding similar sentences across multiple languages in wikipedia. *Proceedings of the 11th Conference of the European. Chapter of the Association for Computational Linguistics*, 2007.
- [18] Krizhanovsky, Andrew. Synonym search in wikipedia: Synarcher. *11-th International Conference "Speech and Computer" SPECOM'2006. Russia, St. Petersburg, June 25-29*, pp. 474-477.
- [19] Fernanda B. Viegas, Martin Wattenberg, Kushal Dave. Studying cooperation and conflict between authors with history flow visualization. *In Proceedings of the 2004 conference on Human factors in computing systems*.

- [20] Jakob Voss. Measuring wikipedia. *In Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005.
- [21] Rodrigo Almeida, Barzan Mozafari, Junghoo Cho. On the evolution of wikipedia. *Proceedings of the Second International Conference on Weblogs and Social Media*, 2007.
- [22] Hans Geser. From printed to “wikified” encyclopedias:sociological aspects of an incipient cultural revolution. Technical report, Sociology at the University of Zurich, 2007.
- [23] Bellomi Francesco, Bonato Roberto. Network analysis for wikipedia. Network analysis for Wikipedia. Proceedings of Wikimania 2005 The First International Wikimedia Conference. Frankfurt , Germany., 2006.
- [24] Torsten Zesch, Iryna Gurevych. Analysis of the wikipedia category graph for nlp applications. *in Proceedings of the Workshop TextGraphs-2: Graph-Based Algorithms for Natural Language Processing at HLT-NAACL 2007*,pp. 1-8.
- [25] Andrea Cifforilli. Phantom authority , self selective recruitment and retention of members in virtual communities: The case of wikipedia. First Monday , 8(12).
- [26] Fernanda B. Viegas. The visual side of wikipedia. *HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, 2007.
- [27] Todd Holloway, Miran Bozicevic, Katy Borner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity* , Vol.12 , No. 3,pp. 20-40, 2007.
- [28] Felipe Ortega, Jesus M. Gonzalez Barahona. Quantitative analysis of thewikipedia community of users. *Proceedings of the 2007 international symposium on Wikis*,pp.75-86.
- [29] Yoshiaki Arai, Tomohiro Fukuhara, Hidetaka Masuda, Hiroshi Nakagawa. Analyzing interlanguage links of wikipedias. Wikimania 2008 Conference. Alexandria , Egypt.

## A 付録

表 7: 日本語版 Wikipedia の上位 2 階層のカテゴリ

| 階層 | カテゴリ名                                |
|----|--------------------------------------|
| 0  | 主要カテゴリ                               |
| 1  | 総記, 学問, 技術, 自然<br>社会, 地理, 人間, 文化, 歴史 |

表 8: 英語版 Wikipedia の上位 2 階層のカテゴリ

| 階層 | カテゴリ名   |
|----|---|
| 0  | Contents  |
| 1  | Articles , Categories , Wikipedia featured content<br>Image galleries , , Glossaries , Lists , Portals<br>Timelines , WikiProjects , Wikipedia help<br>Wikipedia administration , Wikipedians |

表 9: カテゴリ ‘人間’ のサブカテゴリ例 (階層別)

| 階層 | 項目名  |
|----|--|
| 1  | 人間   |
| 2  | 生活, 人間科学, 人物, 人体, 健康   |
| 3  | 食文化, 看護学, 大陸別の人物, 人体のしくみ, 健康法  |
| 4  | 政治家, 医学, 宗教, 教育学, 高齢者  |
| 5  | 人権侵害, 小児科学, 日本人の姓, 日本の人物, 靴  |
| 6  | ホラーの登場人物, 剣道漫画, ゴルファー,<br>埼玉県出身の人物, 中国の歌手  |
| 7  | 火の鳥, 日本のプロ野球界, ユタ州上院議員,<br>高知県知事, 日本の司会者   |
| 8  | NEC レッドロケットの選手, JBL の選手,<br>社会人野球, ワシントン・レッドスキンの選手,<br>バレーボールアジア選手権<br>アイシンシーホースの選手, |
| 9  | MLB オールスターゲーム MVP,<br>リンク栃木ブレックスの選手,<br>レラカムイ北海道の選手,<br>パナソニックトライアンスの選手              |
| 10 | 該当なし   |

表 10: イルカに関する言語別トピック (日中英韓)

| 言語  | 関連語   |
|-----|---|
| 日本語 | イルカ水族館, クジラ, イルカ交通<br>どこイルカ, イルカ イラスト<br>イルカと泳ぐ, イルカ 写真<br>白イルカ, イルカナゴり雪, イルカ料理 |
| 中国語 | なし  |
| 韓国語 | ピンクのイルカ, イルカの写真, イルカの音<br>イルカのオリンピック, イルカの絵<br>ラワヂイルカ, イルカのチューブ                 |
| 英語  | イルカの写真, サメ, ドルフィン・フィッシュ<br>イルカゲーム, イルカ情報, バンドウイルカ<br>切り札, シロナガスクジラ              |