2D3-02

Wikipediaの言語間リンクに関する分析

Analyzing interlanguage links of Wikipedia

新井　嘉章[*1]　福原　知宏[*2]　増田　英孝[*1]　中川　裕志[*3]
Yoshiaki Arai　Tomohiro Fukuhara　Hidetaka Masuda　Hiroshi Nakagawa

[*1]東京電機大学 未来科学部
School of Science and Techonology for Future Life, Tokyo Denki University

[*2]東京大学 人工物工学研究センター
Reseach into Artifacts Center for Engineering, The Univer-sity of Tokyo

[*3]東京大学 情報基盤センター
Information Technology Center, The University of Tokyo

Interlanguage-links (ILLs) among Wikipedias are one of important multilingual resources. In this paper, we describe (1) an analysis results of ILLs in Chinese, Japanese, Korean, and English (CJKE) Wikipedias, (2) evaluation results of ILLs using some bilingual dictionaries, and(3) our cross-lingual keyword navigation system using ILLs.

1. はじめに

ウェブ検索に用いるキーワードを他の言語に対訳するとき，既存の対訳辞書を用いて訳が引けない事がある。例えば,「薔薇のない花屋」,「ごくせん」,「あいのり」のようなテレビ番組名,「中川翔子」,「上地雄輔」,「関ジャニ∞」などの人名は,「Yahoo!検索ランキング」に登録されているが,「英辞郎 on the Web」には登録されていない。また,「バスケットボール大韓民国代表」のような，複数の見出し語で構成されるキーワードは，単語に分割しなければ訳が引けない場合がある。

本研究では，検索キーワードの対訳に向けて，Wikipedia[1]を活用した多言語対訳辞書を構築する。Wikipedia は，現在，ドイツ語，フランス語，ポーランド語，日本語，オランダ語，イタリア語，ポルトガル語，スペイン語，スウェーデン語など，255 言語で展開されているウェブベースの百科事典であり，現在，様々な分野への応用が期待されている。中山ら [2] はシソーラス辞書構築の為の言語資源として Wikipedia を活用した。同氏らによれば，自然言語処理などによる従来の構築手法には,「辞書作成時と利用時のタイムラグにより最新の語や概念への対応が困難である」という問題があるとされており，Wikipedia に対する Web マイニングによる辞書構築手法が提案されている。また，Wikipedia 自体の研究も活発に行われている。Ortega[3] らは，10 言語の Wikipedia(英，独，仏, 日，オランダ語，イタリア語，ポルトガル語，ポーランド語，スウェーデン語，スペイン語) について，Gini 係数による投稿者数と投稿記事数の関係の分析を行い，全ての言語で，Gini 係数 0.9 以上 (少数の大量記事投稿者と，多数の少数記事投稿者) である事を確認した。このように，Wikipedia は豊富な言語資源である一方，やや偏った背景を有している。

Wikipedia の特徴の一つとして言語間リンクがある。Wikipedia の各項目の著者や編者は，他の言語版 Wikipedia の同一項目へのリンクを設定できる。例えば,「あいのり」という項目には，英語版への「Ainori」，中国語への「戀愛巴士」などのリンクが設定されている。本研究では，それらのリンク情報から訳語抽出を行うことで多言語対訳辞書を構築する。

本稿では，Wikipedia の言語間リンクの接続パタンと，各接続パタンの割合を示し，言語間リンク情報から得た訳語を一般辞書を用いて評価した結果について述べる。また，言語間リンク情報に基づくキーワードの多言語対訳システムについて紹介する。本稿の構成は次の通りである。2. では，言語間リンクの分析結果を述べる。3. では，評価結果を述べる。4. では，言語間リンク情報に基づくキーワードの多言語対訳システムを紹介する。5. では，まとめと今後の課題について述べる。

連絡先: 新井嘉章，東京電機大学未来科学部情報メディア学科，東京都千代田区神田錦町 2 丁目 2 番地，03-5280-3281, ext 2843，03-5280-3592

2. 言語間リンクの接続状態に関する分析

本節では，(1) 使用したデータ，(2) 分析から得られた言語間リンクのパタンについて述べる。

2.1 Wikipedia データ

表 1 に，日本語版 (2007/10/13)，中国語版 (2007/10/14)，韓国語版 (2007/10/11)，英語版 (2007/10/18) から抽出した項目数と全言語対象の言語間リンク数を示す。また，図 1 に 4 言語間における言語間リンク数と各接続の割合を示す。
Geser[4] は，言語間リンク情報の時間的な量の変化を多言語分析したが，我々は, 接続状態に注目した詳細な分析を行う。

2.2 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，図 2 に示す 5 パタンに分類した。分析では，92%が *Pattern C* である。(表 2 参照)。

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。

表 1: 言語間リンクを持つ項目数

	項目数 (括弧内は言語間リンクを持つ項目数)	全言語を対象とする言語間リンク数
英	5,836,167 (895,235 (15%))	4,072,516
日	808,514 (211,390 (26%))	2,050,491
中	352,533 (122,226 (35%))	1,536,757
韓	93,850 (54,797 (58%))	1,061,280

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 (2.08%)	9,200 (2.70%)	317,971 (93.23%)	331 (0.10%)	6446 (1.89%)
日	14,508 (4.79%)	4,697 (1.55%)	278,281 (91.85%)	271 (0.09%)	5,208 (1.72%)
中	16,356 (7.88%)	3,303 (1.59%)	183,958 (88.68%)	1,605 (0.77%)	2,218 (1.07%)
韓	3,517 (2.87%)	661 (0.54%)	114,910 (93.84%)	435 (0.36%)	2,934 (2.40%)

我々は次の手法によって，約 1%(約 9 千語) 対訳抽出率を向上した。まず，*Pattern A* を接続先によって，*Pattern A-1* と *Pattern A-2* の 2 つに分類する (図 3 参照)。*Pattern A-1* は，接続先にリダイレクト設定がある場合であり，言語間リンクを持つ他の項目へ接続する可能性がある。このパタンは，*Pattern A* の 26%である。更に，*Pattern A-1* を *Pattern A-1-1* から *Pattern A-1-4* に展開すると，*Pattern A-1-4* の間接的な相互リンクの状態があり，この場合，互いに対訳抽出可能である。*Pattern A-1* の内 44%が *Pattern A-1-4* である。よって，*Pattern A* の約 11%(26%中の 44%) は互いに対訳抽出可能となる。

次に, *Pattern B* についても分析し, 約 31% が *Pattern B-1* に分類され (図 4 参照)，*Pattern B-1-2* の間接的な相互リンクの割合は 89%であった。よって, *Pattern B* の約 28%(31%中の 89%) は，間接的な相互リンクとして対訳抽出が可能である。

本節では，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。また，相互リンクでないパタンについても，リダイレクト情報を用いた間接的な相互リンクへの可能性がある事を示した。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，言語を越えた項目間の結び付きに関しては信頼性が非常に高く，対訳辞書構築に応用できると言える。

3. 言語間リンクの内容の分析

相互リンクの言語間リンクを持つ Wikipedia の項目を無作為に 200 件選び，既存の対訳辞書を用いて評価した。

3.1 評価に用いた辞書

日 ⇔ 英 英辞郎 on the Web

日 ⇔ 中 小学館中国語デジタルマルチ辞典 version2.0

日 ⇔ 韓 小学館中韓・日韓辞典 version2.0

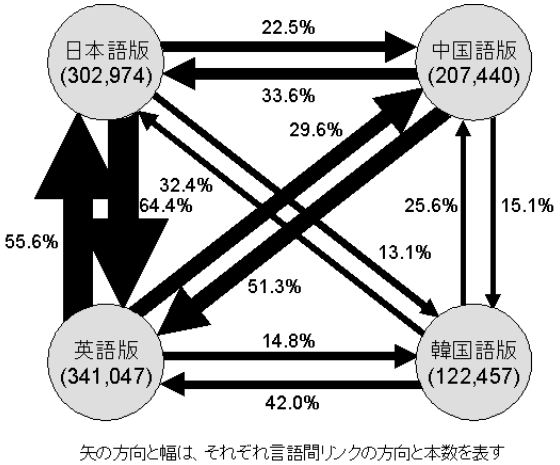

図 1: 4 言語間の接続割合 (ノード内の数値は 4 言語に限定した場合の各言語のリンク数)

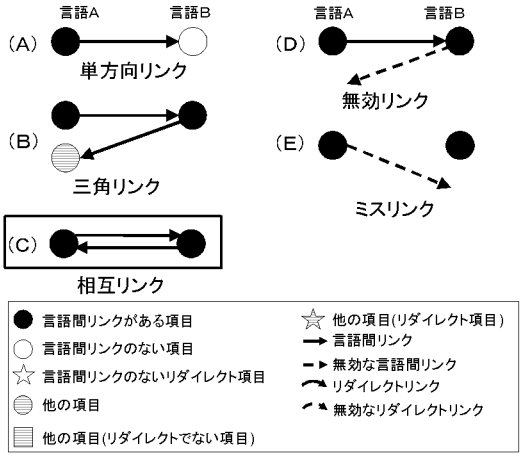

図 2: 言語間リンクの基本パターン

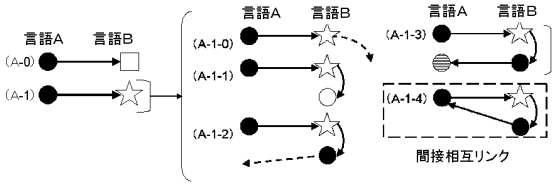

図 3: *Pattern A* の内訳

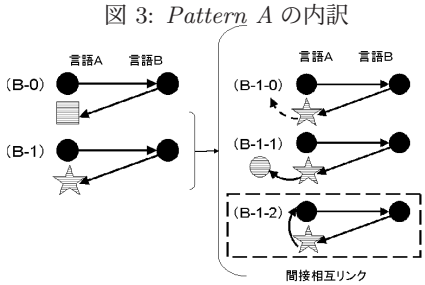

図 4: *Pattern B* の内訳

表 3: 対訳辞書を用いた訳語の評価

	辞書未登録語	辞書と訳語が一致	辞書と訳語が不一致
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%)

表 4: 辞書未登録語の内訳

	部分対訳不可のもの	部分対訳可能なもの
日 ⇔ 英	96 (64.4%)	53 (35.6%)
日 ⇔ 中	120 (70.6%)	50 (29.4%)
中 ⇔ 韓	129 (68.3%)	60 (31.7%)

3.2 評価結果

言語間リンクが設定されている Wikipedia の項目名が，評価用辞書の見出し語に登録されている場合について，評価用辞書による訳語と言語間リンクによる訳語を比較した。その結果として，訳語が一致した項目数を，表 3 の「辞書と訳語が一致」の欄に示し，訳語が一致しなかった項目数を「辞書と訳語が不一致」の欄に示す。また，言語間リンクが設定されている Wikipedia の項目名が，評価用辞書の見出し語として登録されていない場合について，その項目数を「辞書未登録語」の欄に示す。「辞書と訳語が一致」,「辞書と訳語が不一致」,「辞書未登録語」の各パタン毎の詳細を以下に示す。

3.2.1 辞書と訳語が一致

200 件のうち，約 12%は Wikipedia の項目名と既存の対訳辞書の見出語が一致し，訳語も一致した。例として,「音程」,「ナパーム弾」,「意思決定」などがある。

3.2.2 辞書と訳語が不一致

200 件のうち，約 3%は，Wikipedia の項目名と既存の対訳辞書の見出語が一致し，訳語が一致しなかった。例として,「女性専用車両」,「ギャル」,「オード川」などがある。また,「女性専用車両」の場合，言語間リンクによる訳語は「Women-only passenger car」であるのに対し，既存の対訳辞書による訳語は「women-only car」,「women-only train car」,「women-only train coach」のように異なる。この場合，言語間リンクによる訳語は正しく，辞書に無い訳語候補が獲得できた例である。

3.2.3 辞書未登録語

評価した 200 件のうち，約 85%は Wikipedia の項目が辞書の見出語として登録されていない語（辞書未登録語）であり，訳語の比較が出来ない。しかし，それらの項目名の中には，既存の対訳辞書の見出語で構成されているもの（部分対訳可能なもの）が含まれる。表 4 に，部分対訳不可能なものと，部分対訳可能なものの割合を示す。表より，辞書未登録語のうち，約 68%は部分対訳不可能な語であり，約 32%は部分対訳可能な語である事がわかる。前者には,「オダギリジョー」,「多賀町」,「イーエスコウ城」などが含まれる。分類方法として，例えば,「イーエスコウ城」の場合，構成要素である「イーエスコウ」,「城」に分割すると,「城」は既存の対訳辞書の見出語にあるが,「イーエスコウ」は見出語に無い為，部分対訳不可能に分類する。部分対訳可能な語には,「サッカーバハマ代表」,「NBC ニュース」,「ナンディ国際空港」などがある。これらは，構成要素のすべてが，既存の対訳辞書の見出語である為，部分対訳可能に分類する。

3.3 考察

言語間リンクから抽出した訳語は，辞書に未登録の語で約 85%を占め，既存の対訳辞書による比較が出来ない。一方，辞書に登録されていない訳語も得ることができる。

4. キーワードの多言語対訳システム

我々は,Wikipedia の言語間リンクを可視化するシステムを構築した。システムは次の機能を有する。

4.1 訳語の提示機能

Wikipedia の言語間リンク情報に基づいて，入力語に関する他の言語の訳語候補を利用者に提示する。

4.2 カテゴリの提示機能

Wikipedia のカテゴリ情報に基づいて，入力語が属するカテゴリを利用者に提示する。

4.3 関連語の提示機能

Wikipedia のリダイレクト情報に基づいて，入力語に関連するキーワードを利用者に提示する。

4.4 リンク情報の可視化機能

言語間リンク情報とカテゴリ情報を可視化する。クリッカブルマップで表示の視点を変更できる。

4.5 訳語の可視化例

図 5 の例では,「コンピュータゲーム」を入力語として，各言語への対訳を可視化している。また，他の言語の言語間リンクを辿る事で,「コンピュータゲーム」の異表記である「コンピューターゲーム」や，同義語である「パソコンゲーム」が得られた。また，英語の対訳である「Computer and video games」の同義語として,「Video game」と「Personal computer game」が得られた。

4.6 カテゴリの可視化例

図 6 の例では，入力語「ブドウ」のカテゴリ「ワイン」,「果物」,「ブドウ科」,「つる植物」を矩形ノードで可視化した。「ワイン」,「果物」には言語間リンク情報があるのに対し,「ブドウ科」,「つる植物」には言語間リンク情報がない事がわかる。言語間リンク情報がある「ワイン」,「果物」のカテゴリに関しては，英語の「Wine」や「Fruit」といった，訳語にあたるカテゴリを提示できる。このように，言語間でカテゴリを往来可能になる為，今後，言語間リンク情報のない項目に対して，言語間リンクを自動補完する仕組みへの応用が期待できる。

5. おわりに

本稿では，言語間リンクの接続状態と，各接続状態の割合を示した。言語間リンクは，約 92%が相互リンクの状態であり，言語を越えた項目間の繋がりに，信頼性が認められた。更に，相互リンクでないパタンについて，リダイレクト経由での対訳抽出の可能性を示した。評価として，既存の対訳辞書を用い，辞書未登録の語で約 85%占めている事を明らかにした。また，Wikipedia の言語間リンク情報を用いた対訳システムについて述べ，言語毎のカテゴリ間の関係性や，多言語での同義語や異表記語の発見が可能になる事を述べた。今後の課題として，既存の対訳辞書を用いて訳の正しさを確認できない訳語についての評価が挙げられる。

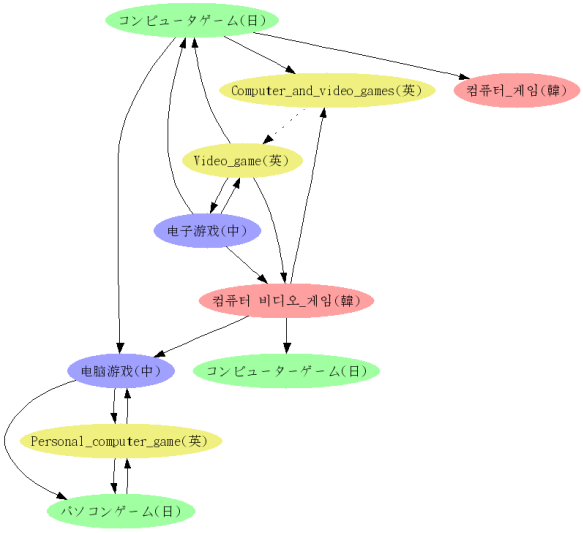

図 5: 言語間リンクの可視化

参考文献

[1] Wikipedia. フリー百科事典『ウィキペディア』. http://www.wikipedia.org/.

[2] 中山浩太郎, 原隆浩, 西尾章治郎. Wikipedia マイニングによるシソーラス辞書の構築手法. 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928, 20061015.

[3] Felipe Ortega, Jesus M. Gonzalez-Barahona, Gregorio Robles. On the inequality of contributions to wikipedia. *Proceedings of the 41st Hawaiian International Conference on System Sciences (HICSS-2008), 2007.*, 2007.

[4] Hans Geser. From printed to "wikified" encyclopedias:sociological aspects of an incipient cultural revolution, 2007. (Available at online, http://socio.ch/intcom/t_hgeser16.htm).

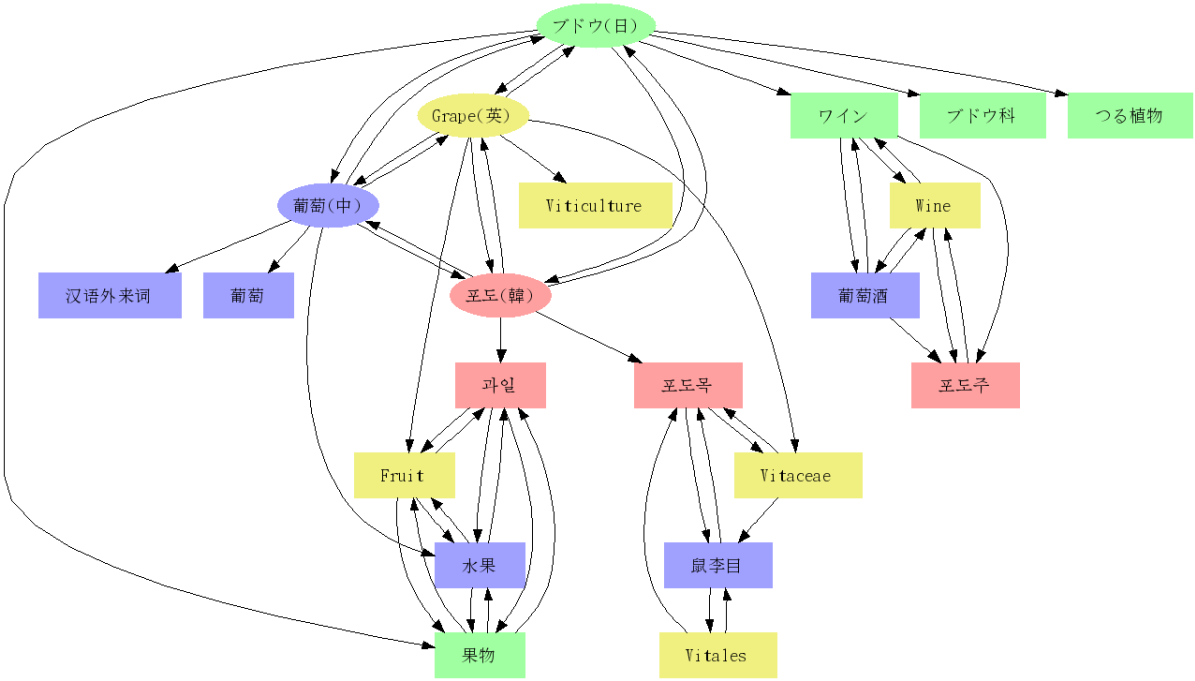

図 6: カテゴリ情報の可視化