

Wikipediaを用いた多言語ブログ検索のための訳語抽出

新井 嘉章[†] 福原 知宏^{††} 増田 英孝[†] 中川 裕志[‡]

[†]東京電機大学 未来科学部 ^{††}東京大学 人工物工学研究センター [‡]東京大学 情報基盤センター

1 はじめに

我々は、多言語ブログ検索システムの開発を行っており、その為の言語資源として Wikipedia[1] を用いる。Wikipedia の特徴の一つに言語間リンクがあり、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。本研究では、言語間リンクから多言語対訳辞書を構築する。

本稿では、我々が Wikipedia の言語間リンクを分析して得た接続パターンを示し、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。

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報告する。3では、キーワードの多言語対訳システムを紹介する。4では、まとめと今後の課題について述べる。

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ーンについて述べる。

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および全言語を対象とする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。

尚、Geser[3] は言語間リンクの時系列変化に注目して分析しているが、本研究では、言語間リンクの接続状態に注目して分析を行う。

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 パターンに分類した。我々の分析では、92%が *Pattern C* の互いに対訳抽出可能なパターンに分類された (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。

Extracting Word Translations from Wikipedias for Multilingual Blog Search

[†] Yoshiaki Arai

^{††} Tomohiro Fukuhara

[†] Hidetaka Masuda

[‡] Hiroshi Nakagawa

School of Science and Technology for Future Life, Tokyo Denki University ([†])

Research into Artifacts Center for Engineering, The University of Tokyo (^{††})

Information Technology Center, The University of Tokyo ([‡])

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

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。

我々は次の手法によって、約 1%(約 9 千語) 対訳抽出率を改善した。まず、*Pattern A* を 2 つに分類する (図 3 参照)。 *Pattern A-1* にはリダイレクト設定があり、言語間リンクを持つ他の項目に接続する可能性がある。このパターンは *Pattern A* の 26% である。その中で、*Pattern A-1-4* は間接的な相互リンクの状態にあり、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。

Pattern B については、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%) は互いに対訳抽出が可能になる。

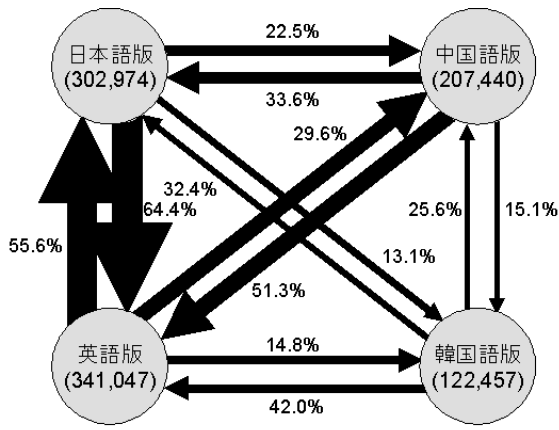
このように、言語間リンクの多くは互いに対訳抽出が可能であり、多言語対訳辞書の構築が可能である。

3 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。利用者は、図 5 のように同義語や他の言語の対訳候補を視覚的に確認可能である。例では、

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%)



矢の方向と幅は、それぞれ言語間リンクの方向と本数を表す

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)

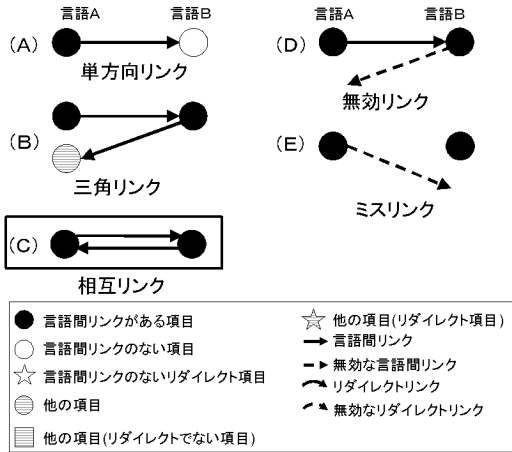


図 2: 言語間リンクの基本パターン

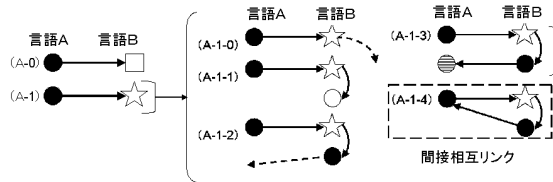


図 3: Pattern A の内訳

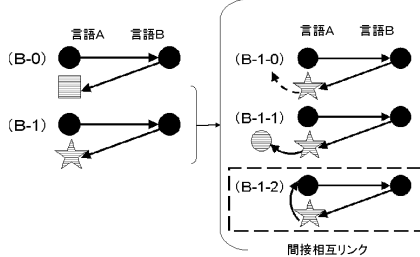
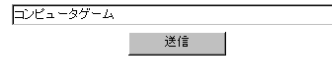


図 4: Pattern B の内訳

(日本語版)



● 類義語: (コンピュータゲーム,ビデオ・ゲーム)

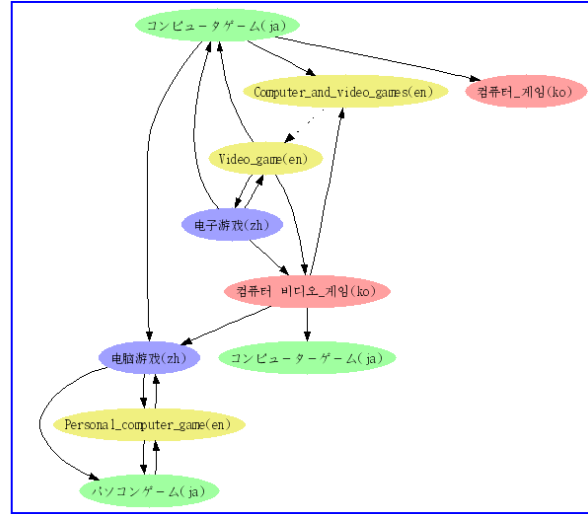


図 5: 言語間リンクの可視化

‘コンピュータゲーム’を入力語として、各言語への対訳と、同義語として‘パソコンゲーム’が、また英語の同義語として‘Video_game’が得られた。

4 おわりに

本稿では、言語間リンクの接続状態と各割合を示した。今後の課題として、実際の検索キーワードを言語間リンクで対訳する調査があげられる。

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] Wikimedia. Wikipedia:全言語版の統計.
http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[3] 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).