多言語に展開するWikipediaの特性調査

森 竜也* 増田 英孝† 中川 裕志‡ 清田 陽司§

概 要

本研究では Web 百科事典 Wikipedia の多言語展開に着目し，それぞれの言語版でどのような特性や差異があるか調査する．Wikipedia は通常の百科事典と異なり，多数の言語版が独立・並行して編集されている．その結果，言語によって Wikipeida の発展の傾向に差異が生じている．このことは問題を含んでいる一方で，各言語版 Wikipedia の特徴を取得するための有用な情報を含んだ興味深いことでもある．本稿ではそのような特性や差異を調査する手法と，研究の経過を報告する．

1 はじめに

Wikipedia[1] は 2001 年に始まった Web 上のフリー百科事典である．Wikipedia の記事はユーザ自身によって作成・編集されていて，日々成長を遂げている．2009 年 9 月現在，日本語版で 60 万件，英語版では 300 万件を超える記事が存在している．これは従来の紙媒体の百科事典を遥かにしのぐ数である．Wikipedia の記事の主題は一定の規則のもとに自由であり，非常に豊富な話題の記事が存在する．またハイパーリンクによる記事同士の参照という Web 特有の情報は，記事と記事，概念と概念のつながりを示す有用な情報である．記事のリンクには同じ言語版の記事を参照するリンクのほかに，250 を超える異なる言語版の Wikipedia の同一概念へのナビゲーションである言語間リンクがある．例えば日本語版の記事 “自然言語処理” には言語間リンクとして，英語版の “Natural language processing” の他にもドイツ語，フランス語，中国語，アラビア語などの言語版の記事へのリンクが存在する．

従来の Wikipedia 研究において，言語間リンクは主に記事のタイトルの翻訳語を獲得するために用いられてきた．Erdmann らの研究 [2] は言語リンクを起点として翻訳語を収集し対訳辞書を自動構築するものである．Auer らによって開発されている DBpedia[3] はセマンティック Web のための知識ベースを Wikipedia から生成する試みで，翻訳語を得るために言語間リンクを利用している．

しかしそれぞれの言語版の Wikipedia は独立・並行して編集されていて，それらの間には差異が生じている．英語版の記事数は日本語版のおよそ 5 倍だが，日本語版のすべての記事を英語版が持っているわけではない．日本語版のみに存在する記事や，逆に他の大部分の言語版では存在するが日本語版にだけ存在しない記事がある．そのような記事の偏りと Wikipedia 内の記事の分類構造を結び付けることで，各言語版 Wikipedia の特性を調査するのが本研究の目的である．研究の成果として，ある言語版で着目されている分野を特定したり，逆に他の言語版に比べて整備されていない分野を見つけることを目標としている．またそういった分野のうち，特に専門的で重要性の高い記事を特定する狙いもある．

*東京電機大学大学院未来科学研究科コミュニケーションデザイン研究室 mori @ cdl.im.dendai.ac.jp
†東京電機大学大学院未来科学研究科コミュニケーションデザイン研究室 masuda @ cdl.im.dendai.ac.jp
‡東京大学情報基盤センター図書館電子化研究部門 n3 @ r.dl.itc.u-tokyo.ac.jp
§東京大学情報基盤センター図書館電子化研究部門 kiyota @ r.dl.itc.u-tokyo.ac.jp

2 Wikipediaのデータ収集

Wikipedia では全データが XML ファイルとしてダウンロード可能になっている．このファイルを解析して他の言語版との比較を行うためのデータを収集する．本研究では次のような Wikipedia 内の情報を収集し，利用する．

カテゴリ

Wikipedia では記事を分類するために自由な文字列によるカテゴリを付与できる．カテゴリにカテゴリを付与することも可能で，記事とカテゴリによる階層的な分類構造が構築されている．記事とカテゴリをまとめてエントリという．1 つエントリに複数のカテゴリを付与することも可能である．ここではどのエントリがどのカテゴリに属するか，あらかじめ全て記録しておく．

バックワードリンク

記事の文章中に他の記事へのハイパーリンクを記述できる．ある記事に着目したとき，その記事へ向けられているリンクがバックワードリンクである．ここではどのエントリが，どのエントリからのバックワードリンクを持っているか調べる．

言語間リンク

異なる言語版の Wikipedia の同一概念へのリンクが言語間リンクである．ここではどのエントリが，どの言語版への言語間リンクを持っているか調べる．

図 1 はこれらのデータの例である．“自然言語” にとっては “計算言語学” のからのリンクがバックワードリンクとなる．またそれぞれがカテゴリ “言語学” に属している．“自然言語” と “Natural language” は日英で相互に言語間リンクが存在する．またカテゴリである “言語学” と “Linguistics” にも相互の言語間リンクが存在する．

これらのデータを XML ファイルから抽出し，全文検索エンジン Lucene[4] のインデックスとして記録しておく．XML ファイルの形式は言語版によらず形式が統一されているので，全て同じ手法でインデックスを生成できる．

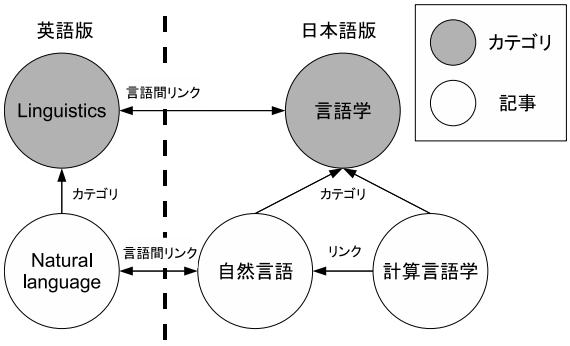

図 1: Wikipedia の構造例

3 特徴データの抽出

言語版を特徴付けるデータにはいくつもの種類が考えられる．本研究で抽出した特徴データと，そのデータが持つであろう意味合いを説明する．

記事のバックワードリンク数

記事のバックワードリンク数はその記事が別の記事から参照されている回数である．文章中で出現頻度の高い一般的な語や，有名な事柄の記事のバックワードリンク数は多く，特定の分野内でしか出現しなかったり，マイナーな事柄の記事のバックワードリンク数は少ないと予想する．

記事の近傍からのバックワードリンクの割合

バックワードリンクのうち，カテゴリ構造内で近傍にある記事からのリンク数を求め，リンク数全体に対する割合を算出する．第 2 章で生成したカテゴリ構造インデックスをグラフ構造として探索し，記事と記事との距離を測る．グラフ構造内においてエントリがノード，カテゴリ付けがエッジとなる．エッジの重みはすべて 1 とする．つまり同じカテゴリに属する記事同士の距離は 2 である．ここでは距離が 2 以下の記事からのバックワードリンクを数える．図 2 は近傍リンクの実例である．

近傍リンク割合は，記事の専門性の指標として用いる．一般語の記事は多くの文脈上に出現するので，カテゴリ構造内の広範な記事から参照され，専門的な記事は近い範囲にある記事から参照されることが

多いと予想する．

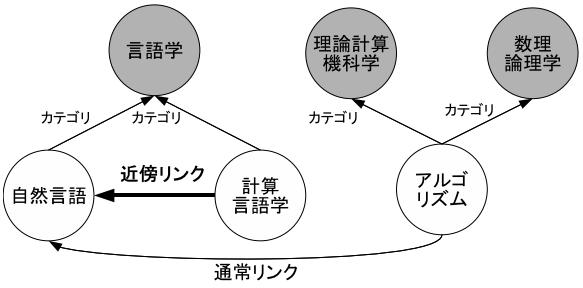

図 2: 近傍リンクの例

言語間リンク数

記事に設定されている言語間リンクの数は，その記事が多言語展開における普遍的な価値の指標と捉えられる．多くの言語版で執筆されている記事は言語間リンクが多く，それだけ普遍的な価値があり，言語間リンクの少ない記事は特定の言語圏，文化圏でのみ話題に上がることが多い記事であると予想する．

4　抽出データの比較

表 1，2 は日本語版および英語版 Wikipedia に対して，第 3 章で述べたデータを抽出した例である．対象とした記事は言語間リンクで接続されたものである．

日本語版 Wikipedia において “日本” のような一般語は参照数が多いが近傍リンクの割合が小さく，カテゴリ構造内の広い範囲から参照されていることが分かる．“自然言語処理” や “形態素解析” のように専門的な語になるにつれ参照数は少なくなるが，近傍リンクの割合が大きくなり，カテゴリ構造内で近い記事からの参照が多いと分かる．本研究ではこのような記事を専門的な記事であると想定している．

英語版 Wikipedia においても日本語版の場合と似た傾向の値が得られた．特に “Natural language” と “Natural language processing” の近傍リンク割合は日本語版と非常に近い．“Text segmentation” と “ChaSen” は参照数が日本語版よりも少ないうえに，近傍リンク数は 0 件であった．後者 2 つは英語よりも日本語の文章の処理において重要な意味を持つ概念の記事である．本研究ではこのような差異を各言語版を特徴づけるものであると想定している．ただし近傍リンク数が 0 件であっても，その言語において専門性がないわけではない．特に参照リンク数が極端に少ない場合，わずかな近傍リンク数の変化が全体に占める割合に大きな変化を与えるので，評価が困難になる．

表に挙げた値はいずれも絶対的な指標となるものではない．参照数はその言語版の持つ記事の総数にも左右される．また近傍リンク割合の値は，他の記事と比べることで大小の評価が可能になるものである．これらの値を活用するにはカテゴリごとに平均を求め，値の大小の評価基準を設定することが考えられる．そうすることでカテゴリに属する記事のうち，有名で重要性が高い記事を判別することができる．

言語間リンク数は普遍的な価値の指標となることを期待して求めたが，この値の大小も評価が困難である．非常に専門性の高い記事は，記事の総数が少ない言語版では存在しない傾向がある．例えば “形態素解析” は “自然言語処理の” 約 6 分の 1 の言語間リンクしか持っていないが，それがそのまま普遍的な価値の比率であるわけではない．また日英で言語間リンクの数に若干の食い違いが見られた．言語間リンクは手動で設定する以外に，整備のためのプログラムが Wikipedia 内で動作しているが，そのプログラムの動作するタイミングや，解析した XML ファイルが生成された日時に差があることが原因と思われる．

5　おわりに

本研究では Wikipedia の多言語展開に着目し，各言語版の特性や特徴を抽出する研究を行った．いくつかの実例では予想に近い結果が得られたが，問題点も明らかとなった．

今後は第 4 章で述べた特徴データの不備を解消した上で，日英以外の版の Wikipedia データに対して

表 1: 日本語版の特性抽出例

記事名	バックワードリンク数	近傍リンク数	近傍リンク割合	言語間リンク数
日本	66978	81	0.001	181
日本語	5375	95	0.018	115
自然言語	131	13	0.099	31
自然言語処理	85	23	0.271	29
形態素解析	36	13	0.361	5
ChaSen	4	3	0.750	1

表 2: 英語版の特性抽出例

記事名	バックワードリンク数	近傍リンク数	近傍リンク割合	言語間リンク数
Japan	74025	36	0.0005	176
Japanese language	6828	33	0.005	110
Natural language	303	30	0.099	31
Natural language processing	261	30	0.241	30
Text segmentation	7	0	0.0	5
ChaSen	2	0	0	1

も同様のデータ処理を行い，各言語版の特徴を浮かび上がらせる予定である．また今回は日英両方に記事が存在する概念について実例を出したが，特定の言語版でのみ存在する，あるいは存在しない記事のうち，重要な記事を求めるという課題への応用を行う．

参考文献

[1] Wikipedia. `http://ja.wikipedia.org/`.

[2] Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Extraction of bilingual terminology from a multilingual web-based encyclopedia. *Journal of Information Processing*, Vol. 16, July, pp. 68–79, July 2008.

[3] Soren Auer, Chris Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, pp. 715–728, 2007.

[4] Lucene. `http://lucene.apache.org/`.