Wikipediaにもとづくキーワード表記ゆれおよび同義語の抽出
Extracting Variant Notations and Synonyms from Wikipedia

石田和成†
Kazunari Ishida†
†広島工業大学　情報学部
† Department of Informatics, Hiroshima Institute of Technology

要旨

Wikipedia におけるエイリアスもとづき,キーワードの表記ゆれ情報の抽出を行った.エイリアスは Wikipedia 内のページ名と異なる名前で, 指定したページへのリンクを張る仕組みである.この仕組みを用いて,Wikipedia の編集者はページ編集において,編集中の文脈に沿ったキーワードを選定して文章を構成する.そのため,本研究ではエイリアスにもとづきキーワードの表記ゆれおよび同義語の抽出を試みた.

1. はじめに

ブログや Wikipedia は個人が情報発信する媒体としてソーシャルメディアを形成している．ソーシャルメディア，特にブログにおいては，様々な個人が個別のブログを持ち情報発信するため，多様なキーワードが用いられる傾向がある．そのため，キーワード量の変化にもとづき社会で話題となっているテーマの調査を行う場合，キーワードの表記ゆれや同義語の取りこぼしにより正確な調査が困難となる．これに対して Wikipedia では，複数の著者によって共有されるページにおいて，個々の概念を説明する文章が整理される．この説明文において他の概念を参照する場合は Wikipedia 内でリンクを作成できる．しかし概念のタイトルが，説明文の文脈に適合しない場合，著者が新たにキーワードを割り当て，エイリアスとして参照できる仕組みが提供されている．この仕組みは，複数の著者が，キーワードの表記ゆれや同義語，関連語の関係を，Wikipedia の文書上に蓄積する仕組みとして利用できる．そのため，この蓄積されたデータにもとづき，表記ゆれや同義語の用語間関係を抽出する．この用語間関係は，キーワードの多様性の高いブログのデータにおいて，流行や意見を抽出するための，キーワードデータベースとして利用できると考えられる．

2. エイリアスにもとづく用語間関係とクラスター抽出

Wikipedia 内の引用で用いられるエイリアスにもとづき，被引用語 t_1 と引用語 t_2，そしてそれらの間の関係 $R(t_1, t_2)$ を抽出する．また，被引用語 t_1 について全被引用数 v_1，引用元語 t_2 について全引用数 v_2，そして，被引用語と引用語の関係 $R(t_1, t_2)$ について関係の出現数 r_{ct} を，それぞれ数え上げる．

表記ゆれ，同義語のクラスターを抽出するために，本研究では被引用語と引用語の関係における到達可能性にもとづきキーワードのクラスターを抽出する．ここで，到達可能性で用いる用語間関係として 4 つの代替案を検討する．1 つ目（無制約）の方法では，データから抽出された全ての関係を用いる．2 つ目（一意）の方法では，引用語が 1 つの被引用語のみと関連がある場合（$v_2=r_{ct}$）を用いる．3 つ目（引用数少）の方法では，引用語の全引用数 v_2 が被引用語の全被引用数 v_1 以下である場合（$v_2<=v_1$）を用いる．4 つ目（被引用数少）の方法では，被引用語の全被引用数 v_1 が，引用語の全引用数 v_2 より小さい場合（$v_1<v_2$）を用いる．

また，到達可能性を調べるにあたり，一方向と双方向の 2 つの代替案を検討する．一方向の場合，一方が被引用語で他方が引用語となるとき，2 つのキーワード間に関係があり，相互に到達可能とする．双方向の場合，双方とも他方の被引用語かつ引用語となるとき，2 つのキーワード間に関係があり，相互に到達可能とする．

Wikipedia のデータは，2009 年 10 月 11 日に生成されたコンテンツのアーカイブ[1] を用いた．このデータから抽出された語数は 3193809 であった．

2.1. クラスターの定量的特徴

抽出方法とクラスターの種類の関連性を調べるため，4 種類の関係性の代替案と，2 種類の方向性の代替案にもとづき，キーワード間の到達可能性によるクラスター抽出を行い，サイズが 2 以上のクラスターに含まれるキーワードを数え上げたところ, 表 1 のような結果が得られた. サイズ 2 以上としたのは, 表記ゆれや同義語のクラスターとして，1 つのキーワードのみのクラスターは意味を持たないからである．表 1 より，一方向の場合，双方向と比べ，多くの単語がサイズ 2 以上のクラスターを形成することが分かった．そのため，Wikipedia におけるクラスター抽出は，一方向を用いることとする．

また，一方向について，得られたクラスターの最大サイズを調べたところ，表 2 に示すように,「2:一意」以外の関係を用いた場合は，非常に大きなクラスターが生じていることが分かった．この大きなクラスターは, 複数のトピックの単語クラスターを含んでおり, 表記ゆれや同義語の抽出に利用できない. そこで，単語数から最大クラスターの単語を除いた単語数を計算したところ,「1:無制約」では 856593,「2:一意」では 964125,「3:引用数少」では 931350,「4:被引用数少」では 229781 となり，最大クラスターサイズが一番小さい「2:一意」において最も多くのキーワードが得られることが分かった.

表 1　語数の比較

	双方向	一方向
1:無制約	10562	1244193
2:一意	3188	964342
3:引用数少	5246	1041290
4:被引用数少	1208	329224

表 2　クラスターの最大サイズ

	語数	クラスター数	最大サイズ	平均語数
1:無制約	1244193	280041	387595	4.442896
2:一意	964342	331050	217	2.91298
3:引用数少	1041290	307078	109940	3.390963
4:被引用数少	329224	64223	99443	5.126263

2.2. クラスターの定性的特徴

クラスターを定性的に調査するために，表記ゆれ，同義語，類義語，同音異義語，異表記異義語の 5 種類にクラスターを分類する．

表記ゆれクラスターは，例えば，2 つの単語を組み合わせた複合語として,「カントリーミュージック」と「カントリー・ミュージック」のように，接続の「・」の有無によるものや，発音のゆれによる「レネー・ゼルウィガー」と「レニー・ゼルウィガー」，アルファベットとカタカナ表記による「Loppi」と「ロッピー」といったものを表記ゆれとする. 表記ゆれクラスターは, 表記ゆれのキーワードのみ含む.

同義語クラスターは，例えば，省略による「ロケテスト」と「ロケーションテスト」，別名による「水玉ブリッジライン」と「水島玉島産業有料道路」といったものを同義語とする．同義語クラスターは，同義語に加えて，表記ゆれを含む場合がある．

類義語クラスターは，例えば，包含関係による「真和中学校・高等学校」と「真和高等学校」，部分文字列による「浦臼町」と「浦臼町営バス」といったものを類義語とする．類義語クラスターは，類義語に加えて，表記ゆれ，同義語を含む場合がある．

同音異義語クラスターは，例えば，動詞について,「振り込む」と「降り込む」，名詞について「鯖」と「椹」といったものを同音異義語とする．同音異義語クラスターは，同音異義語に加えて，表記ゆれを含む場合がある．

異義語クラスターは，例えば,「坂崎」という文字列を含む,「坂崎幸之助のオールナイトニッポン」,「坂崎幸二（のちの幸之助)」,「坂崎一彦」,「坂崎さん」,「坂崎（志摩市)」,「坂崎幸之助」といったクラスターで示されるように，複数のトピックにわたるキーワードが 1 つのクラスターに含まれる．異義

[1] http://download.wikimedia.org/jawiki/20091011/jawiki-20091011-pages-articles.xml.bz2

語クラスターは，異義語に加えて，表記ゆれ，同義語，類義語，同音異義語を含む場合もある．

Wikipedia について，一方向の関係にもとづき到達可能性でクラスターを抽出する．抽出されたクラスターから 100 のサンプルを無作為に抽出し，人手により分類する（表 3）．分類の結果，100 のサンプル中に同音異義語のカテゴリに当てはまるクラスターは観察されなかった．得られた表 3 の結果について独立性の検定を行ったところ，χ^2値は 23.61102 となり危険率 1%で項目間は独立とは言えないことが分かった．表 3 によると，「1:一意」は「表記ゆれ，同義語」が約 8 割，類義語が 2 割であり，異義語はほとんど観察されなかった．そのため，「2：一意」とその他 3 つの方法との間で独立性の検定を行ったところ，「1:無制約」（χ^2値 15.25932）「4：被引用数小」（χ^2値 11.34487）では項目間は独立と言えず関連性があることが分かった．それに対して，「2:一意」と「3:引用数小」は，表記ゆれと同義語を合わせた単語の割合が 79%と 71%であり，前者の割合が高いが，χ^2値が 3.079728 となり統計的に項目間は独立で関連性があるとは言えないことが分かった．しかし「2:一意」は，2.1 において求めた有効語数が最大となるため，用語間関係の条件として用いることとする．表 4 に「2:一意」のクラスターを示す．

表 3　抽出方法とクラスターの種類

	表記ゆれ	同義語	類義語	異義語	合計
1:無制約	11	48	40	1	100
2:一意	18	61	20	1	100
3:引用数小	15	56	29	0	100
4:被引用数小	13	37	48	2	100
合計	57	202	137	4	400

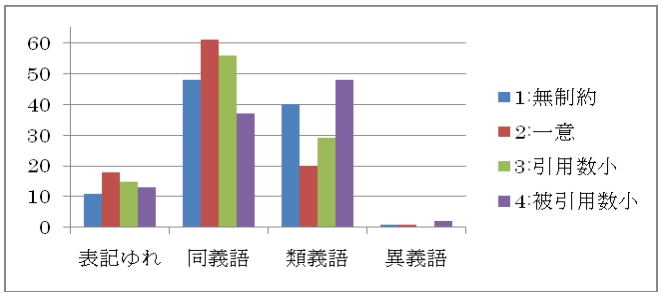

図 1　抽出方法とクラスター分布

表 4　クラスター数とサイズ，単語数

クラスターサイズ	クラスター数	単語数	累積単語数	単語割合
2	224542	449084	449084	0.46568
3	50717	152151	601235	0.62346
4	21798	87192	688427	0.71388
5	11286	56430	744857	0.77239
6	6665	39990	784847	0.81386
7	4136	28952	813799	0.84388
8	2896	23168	836967	0.86791
9	1950	17550	854517	0.8861
10	1425	14250	868767	0.90088
. . . .	. . . .	. . . .	. . . .	. . . .
190	1	190	964135	0.99977
217	1	217	964352	1
合計	331050	964352		

3. 他の情報源との比較

Wikipedia で得られた「表記ゆれ･同義語データ」の特徴を把握するために，他のデータから得られる用語間関係との比較を行う．ここでは，表記ゆれについて人手で編集された辞書である，独立行政法人，国立国語研究所，研究開発部門で作成された「表記統合辞書」[2]を用いる．この辞書は，言語研究・自然言語処理用に開発された，同語判別のための基礎データであり，同音であるが「組み立てる」と「組立てる」のように表記の異なるキーワードをまとめたものである．品詞情報のある見出し語（第 1 フィールド）の語数は 28774，見出し語に対応した語（第 5 フィールド）を含めると，32886 のキーワードが得られた．

3.1. 表記統合辞書におけるクラスターの特徴

Wikipedia の場合と同様に，「表記統合辞書」のデータについて，一方向，双方向それぞれの条件にもとづき到達可能性によりクラスターを抽出した．サイズ 2 以上のクラスターに含まれる語数を調べたところ，一方向の場合が 32886，双方向の場合が 26843（サイズ 1 クラスターに属する単語数は 6043）であった．抽出されたクラスターから，サイズ 2 以上のクラスターから無作為に 100 のクラスターを選定

[2] http://www.kokken.go.jp/lrc/index.php?%A1%D8%C9%BD%B5%AD%C5%FD%B9%E7%BC%AD%BD%F1%A1%D9

し，人手により，表記ゆれ，同義語，関連語，同音異義語，異義語に，それぞれクラスターを分類する．その結果，得られたクラスターは，表記ゆれ，同音異義語，異義語に分類され，同義語，関連語のクラスターは 100 のサンプルにおいては観察されなかった(表 5)．また，一方向は 12%が同音異義語，1%が異義語であったのに対し，双方向では異義語は見つからなかった．これは双方向関係により，語意の異なる関係性が除かれたためと考えられる．2 つの方法と得られたクラスターとの関連性を調べたところ，無関係とは言えないことが確認できた(χ^2値 13.90374)．このサンプリング調査にもとづく語数の推定値は，一方向については，表記ゆれが 28611，同音異義語が 3946，異議語が 329 である．双方向については，表記ゆれが 26843 である．表記ゆれの語数については，一方向の方が双方向と比べて多いが，一方向によるクラスター抽出では 13%のクラスターが意味の異なる単語を含むため，意味の異なる単語のクラスターが観察されなかった双方向の用語間関係にもとづいて抽出されたクラスターのキーワードを Wikipedia のキーワードと比較する．

表 5　表記統合辞書の単語数

	表記ゆれ	同音異義	異義	合計
一方向	87	12	1	100
双方向	100	0	0	100
合計	187	12	1	200

3.2. 考察

表記統合辞書の双方向におけるサイズ 2 以上のクラスターに含まれる単語数は 26843，サンプリングにおいて抽出された用語間関係が全て表記ゆれであった．これはこの辞書が，同音かつ同義かつ異表記の単語間の関係をまとめることを目的としているためである．

それに対して，Wikipedia では，サンプリングの結果，一意，一方向の用語間関係にもとづくサイズ 2 以上のクラスター(全単語数 406908)における推定値は，表記ゆれ 73243(18%)，同義語 2482134(61%)，類義語 81382(20%)，異義語 4069(1%)であった．同義語は，異音かつ同義かつ異表記の単語間関係であり，表記統合辞書では得ることのできない用語間関係である．

また，2 つの情報源から抽出されたサイズ 2 以上のクラスターに含まれる語数について，共通キーワードは 3157, 表記統合辞書の独自キーワードは 23686, Wikipedia の独自キーワードは 958951 であった．つまり Wikipedia と表記統合辞書との間では共通部分は少なく補完的な関係となっていることが分かった．加えて，サンプリングにもとづく人手によるクラスター分類の過程で，Wikipedia は次々と現れる新しいキーワードについての表記ゆれ，同義語の用語間関係が非常に多く含まれることが分かった．

関連する研究として，大島ら[1][2]は，ウェブページのテキストデータから，並列助詞にもとづき，「トマト」「ジャガイモ」といった同位語を発見する手法を提案し，両方向構文パターンにもとづき関連語の高速抽出手法を開発した．また，中山ら[3]，伊藤ら[4]は，Wikipedia のページ間のリンク数やパスの長さで類似度定義を定義し，シソーラス構築手法を開発した．これらの研究に対し，本研究が有利な点は，直接的にエイリアス情報を利用するため，言語の種類に依存せず，計算量も少ない点である．

4. まとめ

Wikipedia におけるエイリアスもとづき, キーワードの表記ゆれ情報の抽出を行うとともに，人手で編集された表記統合辞書との比較を行った．その結果，Wikipedia と表記統合辞書は補完的であり，Wikipedia から抽出されたこの用語間関係は，キーワードの多様性の高いブログのデータにおいて，流行や意見を抽出するための，キーワードデータベースとして利用できる可能性があることが分かった．

参考文献

[1] 大島 裕明,小山 聡,田中 克己,「Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見」,情報処理学会論文誌,データベース,47(SIG_19(TOD_32))　pp.98-1,12, 2006.
[2] 大島 裕明,田中 克己,「両方向構文パターンを用いた Web 検索エンジンからの高速関連語発見手法」,日本データベース学会 Letters Vol.7, No.3, 2008.
[3] 中山 浩太郎,原 隆浩,西尾 章治郎,「大規模 Web 事典からのシソーラス辞書構築」,日本データベース学会 Letters Vol.5, No.4,2007.
[4] 伊藤雅弘,中山浩太郎,原隆弘,西尾章治郎,「Wikipedia からの連想シソーラス構築プロジェクト」,人工知能学会,第 20 回セマンティックウェブとオントロジー研究会,SIG-SWO-A803-05, 2009.