Wikipediaのリダイレクトから得られる同義語の分析

柏岡 秀紀
NICT/ATR
hideki.kashioka@nict.go.jp

1 はじめに

現在の自然言語処理において、表現の多様性は重大な課題の一つとなっている。翻訳処理においては、入力文に意味の多様性を持たせることになり、その訳出も多様性を持つことから、解析、評価が困難になっている。また、Web 上では、同種の情報が複数の異なった表現で記述されている。その膨大な情報にアクセスするためのキーワードが一意に決められず、求める情報を得るために、多様な表現による検索が必要とされる。この表現の多様性を吸収し、翻訳や検索の処理を効率的に行うためには、同義語を収集し、辞書的に管理することが有効である。これまでに、様々なデータから同義語あるいは、言い換え表現を抽出する手法が提案されている [1, 2]。

本稿では、Web 上の知識源として注目されているウィキペディアを利用した同義語の収集について述べ、収集された同義語について分析を加える。

2 同義語の収集

ウィキペディアは、Web 上で自由に利用できる百科事典で、日本語版には、現在 32 万件の記事が掲載されている。基本方針に賛同している人は、記事を投稿、編集することが可能である。日本語だけでなく多数の言語において構築されており、web 上での重要かつ豊富な知識源として広く知られる様になっている。

ウィキペディアの執筆要項には、別称や略称の項目を記述するための特殊なページに関する要項がもうけられている。その一つに、“リダイレクトページ” という項目が設定されている。これは、Web 上の多様なデータへのアクセスを効率的できるように配慮したもので、同義語の収集と同じ目的を持つものである。そのため、現在のウィキペディアに含まれるリダイレクトページの情報を収集し、分析を行った。また、“曖昧さ回避” の項目において、多義性を持つ表現をまとめ、個々のページへのリンクを一覧にしている。

本稿では、ウィキペディアの索引ページから、リダイレクトの記述があると判断した項目について、リンク元とリンク先の項目を同義語として取り出した。

2.1 データ量

ウィキペディアは、日々更新されているため、確定した統計量ではないが、統計として示されている値では、日本語の総項目数は約 32 万記事あるとされている。しかし、ウィキペディアに記述されている索引は完全なものではない。人手による作業により更新されているためである。索引中のリダイレクトは、“トスカナ ⇒ トスカーナ州” のように、記号’⇒’ により結ばれている語のセットと見なして、収集した。ただし、人名については、姓と名の項目名を示すこともあり、リダイレクトされていないこともある。

本稿のデータを取得した時点では、日本語版の索引に記載されている項目が、184,559 あり、リダイレクトされている項目は、36,705 であった。また、曖昧さ回避と思われるものが、7,487 であった。取り出したデータの一部を以下の表 1 に示す[1]。

3 収集データの分類

収集されたデータの中には、リダイレクト元の単語がないもの、曖昧さ回避のためのエントリーと思われるものなどがある。また、対象として同じものを指し示していても、同義語とはいえないものも多く、本来の同義語とは少し変わった扱いとすべきものがある。以下、目についたものを個別に例を挙げて分類する。

曖昧さ回避

曖昧さを回避するためにリンクを張っているもので、“核” という語に対して、“原子核 【 物理学 】、細胞核

[1] ランダムに選択したもので、実際にはリダイレクトではなく、曖昧さ回避のものなども含まれる。

表 1: 取り出されたリダイレクトと思われる対

リダイレクト元	リダイレクト先
関西方言 （かんさいほうげん）	関西弁
キカイダー	人造人間キカイダー
紀元前 613 年 （=-ねん）	紀元前 7 世紀
漁協（ぎょきょう）	漁業協同組合
高知県道 （こうちけんどう）	高知県の県道一覧
昭和通り (埼玉県)	埼玉県道 337 号 久米所沢線
W.B.C. 置換行列 （ちかんぎょうれつ）	世界ボクシング評議会 対称群
バクテリア	真正細菌
ペラン	ジャン・ペラン
メタル Zi	ゾイドジェネシス
18D 形	JR 貨物 18D 形コンテナ
UDF	ユニバーサルディスク フォーマット
Z 基 （-き）	ベンジルオキシカル ボニル基

【 生物学 】、銀河核 【 天文学 】” などが収集されている。リンクの元になる語が収集できていないため、形式的に同義語の抽出からは削除することができる。また、これ以外にも、以下のような、曖昧さ回避の情報も含まれる。

キング 【曖昧】⇒ スティーヴン・キング 、 マーティン・ルーサー・キング、 ドン・キング 、 アーネスト・キング

年代の表現

以下に示す年に関する表記は、ＸＸＸ年をＸ世紀あるいは、ＸＸ０年代へのリンクが作成されている。ウィキペディアのリダイレクトとして記述されている場合と記事中のリンクとして書かれているものがある。

紀元前 247 年 （=-ねん） ⇒ 紀元前 3 世紀
552 年 ⇒ 550 年代
556 年 ⇒ 6 世紀

路線名等交通網に関する表現

以下に示すような路線名などの名称、電車等の名称に関する表現も多く含まれている。

683 系 ⇒ JR 西日本 683 系電車
715 系 ⇒ 国鉄 419 系・715 系電車
伊勢二見鳥羽ライン （いせふたみとば-）⇒ 伊勢二見鳥羽有料道路
稲城大橋 （いなぎおおはし）⇒ 稲城大橋有料道路
茨城県道 57 号常陸那珂港南線 （=ごじゅうななごうひたちなかこうみなみせん）⇒ 常陸那珂道路

組織名・地名

以下のような組織名、地名等においては、略称、歴史的な変遷、通称による表現が含まれる。

大阪外国語学校 （おおさかがいこくごがっこう）⇒ 大阪外国語大学
東大 （とうだい）⇒ 東京大学
東証 （とうしょう）⇒ 東京証券取引所
天神橋六丁目 （てんじんばしろくちょうめ）⇒ 天六

人名

以下に示すように歴史上の人物で別名を持つ、あるいは、表記が異なるものや、芸能人で芸名と本名のような関係のものが含まれる。

姉小路近綱 （あねがこうじちかつな）⇒ 三木近綱
姉小路自綱 （あねがこうじよりつな）⇒ 姉小路頼綱
寺田光男 （てらだみつお）⇒ つんく
ドゴール ⇒ シャルル・ド・ゴール

略称

いわゆる一般の表現で略して表現されるものも、以下のように含まれている。

特番 （とくばん） ⇒ 特別番組
ディスクトップ ⇒ デスクトップパソコン
アニソン ⇒ アニメソング

別称、言い換え

以下のような、別称も多く含まれている。

菜種 （ ナタネ ）⇒ アブラナ
二院制 （にいんせい）⇒ 両院制
2008 年夏季オリンピック （にせんはちねんかき-）⇒ 北京オリンピック
金閣寺 （きんかくじ）⇒ 鹿苑寺
銀閣寺 （ぎんかくじ）⇒ 慈照寺

異表記（表記の揺れを含む）

表記の揺れや異表記による項目も多数含まれている。以下のようなものである。

アーケイックスマイル ⇒ アルカイクスマイル
あずき ⇒ アズキ
カナ ⇒ 仮名（文字）（ かな ）
アーチスト ⇒ アーティスト

4 英語の Wikipedia のデータ

日本語版の Wikipedia でリダイレクト（同義語）を取り出した処理と同様の処理を英語版の Wikipedia に対して行った。(実際には、索引での表記が異なるため、英語版では、リダイレクトの候補を取り出し、各ページがリダイレクトになっているかどうかを確認するという処理を行った。）全体で取り出された項目は、約 12 万件となった。日本語で示した分類と同様の項目が見られる。また、表記として、アルファベットの大文字、小文字のみの差異の項目も多く見受けられた（例：Sf → SF)。頭字語も多く、日本語の略称より項目としては多い様に思われる（例：USA → United States)。また、文字間にスペースがあるだけで、リダイレクトが作成されていることもある（例：O.P. Caylor → O. P. Caylor)。さらに、アルファベット以外の文字の表記なども含まれている。

5 考察

今後、同義語のデータを拡張することを考えた場合、従来は、収集したデータを 3 節で行った分類に従って整理し、その特徴の抽出を試みることが考えられる。地名や組織名等においては、略称が多く、略称の構成規則が推定できれば、現在、取り出せた略称以外に適用して略称の構成が可能と思われる。これまでにも、略称の構成規則を推定する研究は行われているが、短い単語の中の構造を把握する必要があり、また、語や単語の内部構造によって略称として残す文字列が変わることもあり、それほど良い性能が出ているわけではない。そこで、本稿では、Web 上のデータの利用を考え、以下の手順による同義語の抽出を検討する。まず、今回、得られた同義語のペアで Web 検索を行う。次に、検索されたページ内で、両表現を結びつける特徴的な表現を抽出する。そして、新たに同義語を見いだしたい語と、得られた特徴的な表現を利用して、Web を検索し、同義語に相当する語を取り出す。このような特徴的な表現の一つに安直ではあるが、「略称」や、「別名」という語がある。現在、このような特徴を持つ語の洗い出しを行っており、今後、同義語の抽出処理を検討したい。

6 まとめ

本稿では、既存の Web 上の知識を有効に利用することを考え、ウィキペディアに含まれる “リダイレクト” の項目を利用した同義語の収集を行い、固有表現として判別される人名、地名 (場所)、組織名、時間表現、数値などを中心に、その分類を行った。得られた表現は、日本語で約 3 万 6 千対であり、英語では、12 万対となる。ウィキペディアは、他の言語も同様に整備されているため、多言語で同様の同義語対を収集できると考えられる。現在、得られた同義語対を利用した同義語対を拡張して収集する手法を検討している。

参考文献

[1] 乾, 藤田. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[2] 村田, 金丸, 井佐原. 複数の辞書の定義文の照合に基づく同義表現の自動獲得. 自然言語処理, Vol. 11, No. 5, pp. 135–149, 2004.