Wikipediaのエントリ-リダイレクト間を対象にした同義関係抽出

大野 潤一　　柴木 優美　　山本 和英
長岡技術科学大学 電気系
{oono, shibaki, yamamoto}@jnlp.org

1　はじめに

同じ意味を共有する異なる表記の語は同義語と呼ばれ，情報検索，テキストマイニングなどのテキスト処理の分野で有用とされる．大量の日本語の同義語を収録した語彙資源に，日本語 WordNet[1]や ALAGIN 基本的意味関係の事例ベース[2]などがある．これらの語彙資源は人手により高精度な同義語が登録されているが，日々生まれる新語とその同義語を随時辞書に追加するには多大な労力が必要となる．

そこで，これまでに様々なデータから同義語対を自動抽出する研究が行なわれきた．その中に，即時更新性に優れたオンライン百科事典である Wikipedia のリダイレクト機能を利用した同義語対抽出の研究がある．

リダイレクト機能とは，あるページを表示させたときに，別のページに自動的にリダイレクト (転送) する機能のことである．また，そのようなページをリダイレクトページと呼ぶ．リダイレクト先のページは大抵の場合，見出し語と見出し語を説明する文が書かれたエントリページである[3]．例えば図 1 のように，リダイレクトページ “猫じゃらし” はエントリページ “エノコログサ” にリダイレクトされる．リダイレクト名とエントリ名が同義関係にあることに着目し，柏岡 [1] や玉川ら [2] はリダイレクト名-エントリ名を全て同義語対とみなして抽出する手法を提案している．しかし実際にはリダイレクト名とエントリ名は，[ソファ→ いす]，[分骨 → 遺骨] など，同義関係でないことも多く，全てを同義語対とみなすには問題がある．

そこで，本稿ではリダイレクト名とエントリ名のリンクから同義関係を高精度で判定する手法を提案する．リダイレクト名がエントリページの本文中に存在する

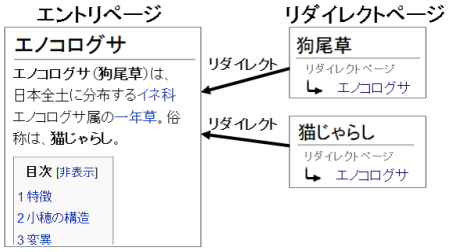

図 1: エントリページとリダイレクトページの例

場合，特定の語彙統語パターンにマッチするリダイレクト名を，エントリ名と同義関係として抽出する．例えば，エントリページの本文中に,「エントリ名（リダイレクト名)」という表記があったとき，エントリ名とリダイレクト名は同義関係である，というようなパターンを作成する．図 1 では，このパターンを用いて「エノコログサ（狗尾草)」という表記から，“狗尾草” と “エノコログサ” を同義関係として抽出する．

2　関連研究

同義関係を自動抽出する手法はこれまで数多く行なわれてきた.「同義語は同じような文脈で使用される」という仮定に基づいて同義語を抽出する研究に酒井ら [3] や下畑ら [4] の研究がある．酒井らは同義語対候補を，文字情報による規則を用いてコーパスから探索し，両者の名詞間類似度が閾値以上だった場合同義関係であるとしている．下畑らは同一言語のパラレルテキストから局所的な文脈一致などの情報をもとに同義語を獲得している．このように文脈情報をもとにした手法は類義関係の高さから判別しているため，例えば対義語との弁別が難しく，適合率が低い．増山ら [5] や小島ら [6] は表層の文字列を利用して同義語関係を抽出する研究をしている．増山ら [5] は，英単語に対応するカタカナの表記ゆれ（例：スパゲッティ，スパゲティー，スパゲテイ) を取得する手法を提案している．小島ら [6] は類似度が高く編集距離が小さい語句対に対し，異表記対かどうかを機械学習を用いて判定している．こ

[1] http://nlpwww.nict.go.jp/wn-ja/
[2] http://alaginrc.nict.go.jp/
[3] 以降，リダイレクトページの見出し語をリダイレクト名，エントリページの見出し語をエントリ名と呼び，リダイレクト名とエントリ名の関係を [リダイレクト名 → エントリ名] といった形で表す．

れらの手法では同義語というよりも同一語の表記ゆれのような場合に適用できる手法であり，全く語句の違う同義語対には適用できない．

岡崎ら [7] や大西ら [8] は語彙統語パターンから同義語を抽出している．岡崎らは「朝鮮民主主義人民共和国（北朝鮮）」のような括弧内とその前の語を同義語対候補とし，機械学習で同義関係を判定している．大西らは国語辞典から「アイス：アイスクリーム の略」(下線部がパターン) のような語彙統語パターンを用いて自動で同義関係を獲得している．

大西らのように国語辞書の定式化された語釈文を用いると，語彙統語パターンのみで高精度で同義語対を抽出できる．しかし国語辞典は新語に弱く，語彙を増やし続けることが難しい．一方ユーザー参加型のオンライン百科事典である Wikipedia は，新語彙の即時更新性に優れている．しかし Wikipedia は自由な形式で書かれているので，語彙統語パターンだけで同義語対を抽出すると精度が低くなる．そこで本手法では，リダイレクト名をエントリ名の同義語候補とし，エントリページ本文中に存在するリダイレクト名が特定の語彙統語パターンにマッチした場合に両者を同義関係として抽出する．

3　リダイレクトページと同義関係

2010 年 11 月 2 日時点の日本語 Wikipedia に対し，リダイレクト名とエントリ名の関係について調査を行なった．同義関係か否かの人手分類基準に，ALAGIN 基本的意味関係の事例ベース Ver1.0[2] を参考にした．この事例ベースは情報通信研究機構で構築された，語句対に対し，同義関係，対義関係，部分・全体関係などを人手で付与したデータである．その際の同義関係の判定基準を本手法で利用した．以下に本手法で使用した同義関係判定の人手分類基準を示す．

同義関係

(a) 同義異語句対：同じ対象を指示する異なる語句の対である場合．
(例) [猫じゃらし → エノコログサ], [県立浜北公園 → 静岡県立森林公園], [ウェブラジオ → インターネットラジオ], [ネズモドキ → ギョリュウバイ]

(b) 略語対：同じ語句の異なる表記の対だが，一方が他方の略式表記になっている場合．
(例)[PMET→ 医療研修推進財団], [大阪駅北 → 大阪駅北地区], [水爆 → 水素爆弾]

(c) 同語異表記対：(b) 略語対以外で，発音 (読み) が同一である語句の対の場合．
(例)[橇 → そり], [トイザラス → トイザらス], [X-box→X-BOX], [うさぎ → ウサギ], [徳川光国 → 徳川光圀], [ジョニーデップ → ジョニー・デップ], [I フォン →IPhone], [Yahoo，Yahoo!]

非同義関係

(例)[要物契約 → 契約], [板野サーカス → 板野一郎], [秋山大地 → ジャニーズ Jr.], [トウジュロ → シュロ], [分骨 → 遺骨]

全てのリダイレクト名-エントリ名のペアから無作為抽出した 2,000 件に対して調査した結果，それぞれの割合は (a) 同義異語句対 18.7%，(b) 略語対 26.7%，(c) 同語異表記対 28.7%，非同義語対 26.0%であった．

4　同義関係抽出手法

リダイレクト名がエントリページの本文中に存在する場合，特定の語彙統語パターンにマッチするリダイレクト名を，エントリ名と同義関係として抽出する．本手法では，エントリページの本文に対し前処理を施したのち，3 種類の語彙統語パターンを用いて同義関係の判定をした．

4.1　前処理

前処理は以下の手順で行なう．

1. エントリページの本文中の第 1 節見出しより下位の文書群を削除する[4]．
2. 本文中の文字列にエントリ名があったら，[entry] に文字列を置き換える．もしエントリ名が囲い記号[5]で囲われていた場合，記号ごと置き換える．
3. 本文中の文字列にリダイレクト名があったら，[redirect] に文字列を置き換える．2. と同様，囲い記号があればまとめて置き換える．
4. 囲い記号で任意の文字列が囲まれていた場合，記号ごと [other] に文字列を置き換える．ただし，任意の文字列中に [entry]，[redirect] がある場合は置き換えない．
5. 「および | または | もしくは | や」のような文節間の並列を表す形態素を，読点に置き換える
6. 括弧内の文字列を抜き出し別の 1 文とする (4.2 節のパターン 3 を適用する場合を除く)．

1. は，誤った同義関係の抽出を防ぐための処理である．高精度で同義関係を抽出するためには，パターンを適用する文がエントリ名について書かれている必要がある．ページの下位に行くほどエントリ名以外の説明が

[4] Wikipedia は MediaWiki 構文で書かれているため，タグを利用して第 1 節見出しを判定できる
[5] 文字列の囲い記号は，‴ 文字列 ‴,「文字列」を指す

書かれることが多いため，第 1 節見出しまでの文章のみに語彙統語パターンを適用する．4. でエントリ名とリダイレクト名以外も文字列を置き換えるのは，“巨人の星” などの名詞句を 1 つの名詞として扱うためである．5. では並列表現を読点に置き換えることで，Wikipedia に多い並列した単語の処理を簡略化する．6. は括弧の中を別の文として語彙統語パターンにマッチさせるための処理である．エントリ名が “エベレスト” でリダイレクト名が “サガルマータ” のときの，前処理前と前処理後の文の例を以下に示す．

＜ 前処理前 ＞
・エベレストのネパールでの名称は ‴ サガルマータ ‴(Sagarmatha) またはサガルマタで，「世界の頂上」という意味である

＜ 前処理後 ＞
・[entry] のネパールでの名称は [redirect]，サガルマタで，[other] という意味である
・Sagarmatha

4.2 同義関係抽出の語彙統語パターン

前処理後の [redirect] を含む文に対し，以下の 3 種類の語彙統語パターンを適用して同義関係を抽出する．

パターン 1:名詞をキーワードとしたパターン

名詞キーワード

自称，通称，呼称，別名，別称，略称，俗称，陵名，名称，公称，愛称，俗称，蔑称，卑称，古称，異称，旧鈔，仮称，院号，和訳，英訳，語訳，日本語訳，前身，改称，改め，名前，名義，呼び名，正式名，四股名，改名，旧名，源氏名，本名，学名，幼名，和名，英名，登録名，芸名，あだ名，渾名，綽名，徒名，仇名，仮名，筆名，登録名，登記上社名，表記，省略，省約，同義，同等の意味，同じ意味，婉曲，遠回し，旧字，題字，略字，元の用字，本来の用字，原題，邦題，ニックネーム，フルネーム，ニックネーム，ハンドルネーム，ペンネーム，ラジオネーム，ペットネーム，タックネーム，TAC ネーム

初めに，上述した名詞キーワードの直前直後の名詞・記号の連続文字列を抽出する．名詞キーワードの前方，後方の形態素を順にチェックしていき，一番初めに出現する名詞・記号の形態素から，次に出現する名詞・記号以外までの形態素の連続を，文字列として抽出する．次にその文字列を読点で区切り，得られた文字列が [redirect] であればエントリ名とリダイレクト名を

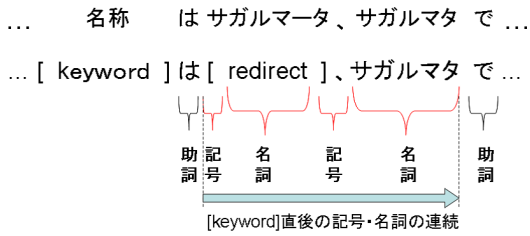

図 2: keyword 直後の名詞・記号の連続文字列の抽出例

同義関係とする．図 2 ではキーワードを “名称” としたとき，連続文字列として “[redirect]，サガルマタ” を抽出する．この文字列をを読点で区切ると [redirect] が得られるため，エントリ名 “エベレスト” とリダイレクト名 “サガルマータ” を同義関係とみなす．

パターン 2:文末表現をキーワードとしたパターン

文末キーワード

呼ばれる，呼ぶこと，称される，称する，しょうされる，しょうする，とも言う，とも言い，とも，ともいう，ともいい，略して，略され，略す，りゃくして，りゃくされ，りゃくす，と命名され，の名で呼ばれる，表現が用いられ，語が用いられ

パターン 1 と同様にして，上述した文末キーワードの直前の名詞・記号の連続文字列を抽出し (直後の文字列は抽出しない)，文字列を読点で区切ったときに得られた文字列が [redirect] であればエントリ名とリダイレクト名を同義関係とする．

パターン 3:括弧表現を利用したパターン

本文中に，「[entry](任意の文字列)」が存在したとき，任意の文字列が [redirect] を含んでいれば，エントリ名とリダイレクト名を同義関係とする．

5 実験と考察

5.1 実験設定

2010 年 11 月 2 日時点での日本語 Wikipedia のダンプデータ[6]を使用して評価実験を行なった．リダイレクト名-エントリ名のリンク数は 406,835 件のうち，リダイレクト名を置き換えた [redirect] がエントリページの本文に存在するリンク数は 156,623 件あった．今回は，全てのリダイレクト名-エントリ名のペアから無作為抽出して人手で正解を付与した 2,000 件に対し，評価を行なった．3 節で同義関係を 3 種類にわけて定義したが，本手法ではそれらを区別しないで抽出した．形態素解析器には MeCab[7]を用いた．

5.2 実験結果

本手法で抽出できるリダイレクト名-エントリ名の同義関係の適合率は 92.1%(151/164) であった．本手法で定義した語彙統語パターンにより正しく抽出された同義関係の例を表 1 に示す．また，同義関係の種類別の再現率を表 2 に示す．表 2 より，無作為抽出した 2,000 件の同義異語句対の総数はペア全体とペアを限定したもののどちらについても略語対，同義異表記対より割合が高く，全体で 14.0%，ペアを限定した再現率で 41.3%だった．よって，本手法は同義異語句対に対して特に有効である．

[6]http://download.wikimedia.org/jawiki/
[7]http://mecab.sourceforge.net/

表 1: 本手法で正しく抽出できだリダイレクト名-エントリ名の同義関係の例

パターン	リダイレクト名 → エントリ名	語彙統語パターンにマッチした文
名詞	獏与太平 → 古海卓二	. . . に脚本を提供したのを最後に,「獏与太平」名義で作品を発表しなく . . .
名詞	ネズモドキ → ギョリュウバイ	針葉樹のネズに似るので '''ネズモドキ''' の別名もある
文末	うだつの町並み → 脇町南町	'''うだつの町並み''' と呼ばれることもある
文末	蘚苔植物 → コケ植物	. . .''' 蘚苔類 '''（せんたいるい），'''蘚苔植物'''（せんたいしょくぶつ）などともいう
括弧	Post-punk→ ポストパンク	'''ポストパンク'''（'''Post-punk'''）は，1970 年代の終わりから勃興した . . .
括弧	神通川第一ダム → 神一ダム	'''神一ダム'''（じんいちダム，'''神通川第一ダム'''）は，発電用ダムである

表 2: 同義関係の種類別の再現率

同義関係種類	ペア全体からみた再現率	ペアを限定 [*1] した再現率
同義異語句対	13.9%(52/373)	41.3%(52/126)
略語対	9.0%(48/533)	28.6%(48/168)
同語異表記対	8.9%(51/574)	39.8%(51/128)
全同義語対	10.2%(151/1480)	35.8%(151/422)

[*1] エントリページの本文中の第 1 節見出しまでに [redirect] が存在するリダイレクト名-エントリ名ペアに限定した再現率

5.3 誤り解析

誤って抽出されたペアには,「特に 〜」,「〜 を除いて」といった限定を表す表記や，エントリの一部のみを説明している文が多かった．再現率が低いのは，語彙統語パターンを網羅しきれていないこと，本文にリダイレクト名がない場合が多く存在するためだと考えられる．再現率を上げるには，語彙統語パターンをさらに増やす必要がある．また，リダイレクト名が本文に存在しない場合でも，エントリ名とリダイレクト名の同義判定を行える手法が必要である．同義関係が判断できない理由の 1 つに，リダイレクト名が本文に存在しなかったということがある．そもそもリダイレクト名は本文の補足のために書かれることが多いので，エントリ名とリダイレクト名の表層の文字列が似ている場合は，リダイレクト名は書かれないことが多かった．

5.4 日本語 WordNet との比較

同義語を数多く収録している日本語 WordNet[1] と，本手法で得られたエントリ名-リダイレクト名の同義語対がどの程度一致しているか調査した．日本語 WordNet では個々の概念はそれぞれ “synset” という単位にまとめられ，同義語や上位語，下位語，関連語などの情報が付与されている．例えば，ある synset には同義語集合として “海豚”, “ドルフィン”,“イルカ” が収録されている．日本語 WordNet(ver.1.0) には 56,741 件の同義語集合が収録されている．同じ同義語集合に本手法で得られた同義語対が存在する件数を調査した結果，同義語対とされた 36,068 件のリダイレクト名-エントリ名ペアのうち，1,172 件 (3.2%) が日本語 WordNet と一致した．一致したペアは [鶺鴒 → セキレイ] や [レシート → 領収書] などの普通名詞であった．Wikipedia を使用したことにより，多くの固有名詞や新語彙の同義関係を抽出することできたと言える．

6 おわりに

本稿では，Wikipedia のリダイレクト名-エントリ名のリンクから，エントリページの本文中に対し語彙統語パターンを適用することで同義関係を抽出した．その結果，適合率 92.1%で，36,068 件の同義関係を抽出することができた．更に Wikipedia を調査して語彙統語パターンを増やし，同義関係抽出の再現率をあげることが今後の課題である．

参考文献

[1] 柏岡秀紀．Wikipedia のリダイレクトから得られる同義語の分析. 言語処理学会第 13 回大会発表論文集, pp.1094-1096, 2007.

[2] 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平. 日本語 Wikipedia からの大規模オントロジー学習. 人工知能学会論文誌, Vol.25, No.5, pp.623-636, 2010.

[3] 酒井浩之, 増山繁．略語とその原型語との対応関係のコーパスからの自動獲得手法の改良. 自然言語処理, Vol.12, No.4 ,pp.207-231, 2005.

[4] 下畑光夫, 隅田英一郎. 単言語パラレルテキストからの同義語獲得. 言語処理学会第 11 回大会発表論文集, pp.1153-1156, 2005.

[5] 増山毅司, 中川裕志. Web データを利用したカタカナ異表記の自動獲得. 言語処理学会第 11 回年次大会, pp.412-415, 2005.

[6] 小島正裕, 村田真樹, 風間淳一, 黒田航, 藤田篤, 荒牧英治, 土田正明, 渡辺靖彦, 鳥澤健太郎．機械学習と種々の素性を用いた編集距離の小さい日本語異表記対の抽出．言語処理学会第 16 回大会発表論文集, pp.928-931, 2010.

[7] 岡崎直観, 石塚満. 言い換え可能な括弧表現の抽出法. 言語処理学会第 13 回大会発表論文集, pp.911-914, 2007.

[8] 大西貴士, 黒橋禎夫．国語辞典からの類義表現抽出と SYNGRAPH データ構造による柔軟マッチング. 言語処理学会第 12 回大会発表論文集, pp.1127-1130, 2006.