課題研究報告書

普通名詞換言辞書の構築

長岡技術科学大学 工学部
電気電子情報工学課程
学籍番号：09108085

山形　祐輝

指導教員 山本 和英 准教授

2014 年 2 月 24 日

概要

語彙の換言は様々な言語処理タスクに関わっており、言語処理において基盤となるタスクである。文を機械的に分析するときに言語は曖昧性に溢れている。その原因として多義性と同義性の問題が挙げられるが、その解決には知識の収集が必要不可欠である。知識の収集方法として、国語辞典の語釈文等の既存の語彙資源から同義関係や上位下位関係などを抽出する方法や、大量の文集合から知識を獲得する方法がある。換言分野では実際に典型的な手法として、これらの言語資源を使った研究は数多くある。しかし、適切な換言候補を高精度で選択することは困難であり、そもそも適切な換言候補が利用する言語資源中に存在しない場合もある。人が実際に行う換言では、上位下位関係や説明的な換言の他に、シソーラスや国語辞典等の語彙資源では獲得出来ない知識も用いて換言を行っていると考えている。

本研究では、形態素解析器 JUMAN の形態素辞書を基に人手で普通名詞換言辞書の構築を行う。JUMAN の形態素辞書に登録されている普通名詞の代表表記を換言対象とし、日本語初学者からその言葉の意味を問われたときにどのように答えるか、を念頭に置いて換言する。JUMAN の形態素辞書には普通名詞が約 1 万 7 千語登録されており、その約 95%にあたる約 1 万 6 千語の換言対を得た。

また、構築した普通名詞換言辞書と我々の研究室で以前構築した用言等換言辞書を合わせた換言辞書の評価として文検索を行う。単語単位のクエリの拡張には同義語、上位語、下位語が有効であることが明らかになっている。比較対象として日本語 WordNet 同義語データベースを用いる。換言辞書と日本語 WordNet 同義語データベースでそれぞれクエリ拡張を行い、元クエリで得られた文と拡張して得られた文で類似度計算を行う。換言辞書と日本語 WordNet 同義語データベース共に獲得文数は約 110,000 文であり、類似度も同程度であった。しかし、元クエリで獲得した文の Jaccard 係数が 0.4 未満となるクエリを足切りした場合、換言辞書による拡張の方が類似度が高くなっており、換言辞書は日本語 WordNet 同義語データベースで拡張を行う場合と同等以上の効果があることがわかった。

目次

1 序論

山本[1]が論じているように、語彙の換言は様々な言語処理タスクにおける基本課題であり、適切な換言処理が可能となれば、各分野の性能向上に大きく貢献すると考えている。たとえば日本語から英語、中国語といった翻訳は異言語間の換言と言える。また要約タスクにおいて、表現をより短くするために同義表現の知識が必要になる。さらに情報検索においてはクエリ拡張に語彙の換言が用いられている。このように、語彙の換言は様々な言語処理タスクに関わっており、言語処理において基盤となるタスクであるといえる。

また、乾[2]は機械的に分析するときに言語は曖昧性にあふれており、その原因として多義性と同義性の問題があると論じている。このような問題の解決には知識の収集が必要不可欠である。その収集方法として、国語辞典の語釈文等の既存の語彙資源から同義関係や上位下位関係などを抽出する方法や、大量の文集合から知識を獲得する方法がある。換言の分野では実際に典型的な手法として、これらの言語資源を使って換言知識を獲得する研究は数多くある。しかし、適切な換言候補を高精度で選択することが困難であり、そもそも適切な換言候補が利用する言語資源中に存在しない場合もある。人が実際に行う換言では、上位下位関係や説明的な換言も行っているが、シソーラスや国語辞典等の語彙資源では獲得出来ない知識も用いて換言を行っていると考えている。

そこで、本研究では普通名詞に対して完全に手作業で換言辞書を構築する。同様の目的で、山本ら[3]は用言等換言辞書として動詞、サ変名詞、形容詞、副詞についての換言辞書を構築している。

さらに、構築した普通名詞換言辞書と用言等換言辞書を合わせた換言辞書をクエリ拡張に用いることで有用性を示す。

2　関連研究

換言辞書に類似した言語資源として国語辞典や、日本語 WordNet [4]や日本語語彙大系等のシソーラスが挙げられる。また、このような言語資源を構築する研究は数多くある。柴木ら[5]は Wikipedia の記事に付与されているカテゴリの階層構造を利用して is-a 関係オントロジーを構築している。ここで is-a 関係とは” B is a A ”となるような上位-下位関係のことである。また、難波ら[6]は特許データベースのテキストから「などの」、「等の」という 2 つの定型表現を用い、シソーラスを自動で構築している。他にもこのような言語資源を自動で構築する研究はある[7]が、人による換言の知識が得られる汎用な換言辞書というものは作られていない。

換言に関する研究において、国語辞典やシソーラス等の言語資源を換言知識として用いることは典型的な手法である。梶原ら[8]は国語辞典の語釈文中の語を換言候補として、シソーラス中の距離計算に基づいた換言手法を提案している。呉[9]は国語辞典やシソーラスの他、コーパス、WEB 等の言語資源を用いて語彙、構文、意味の三つの段階における言い換え知識の獲得に有効なパターンを提案している。しかし、このような既存のシソーラス等を用いた手法の場合、1 章でも述べたように適切な換言候補を高精度で選択することが困難であり、そもそも適切な換言候補が利用する言語資源中に存在しない場合もある。

このような背景から、普通名詞について完全な人手による換言辞書の構築を行う。また、その評価としてクエリ拡張を行う。

Ellen M ら[10]は情報検索において、検索初期段階の洗練されていない単語単位のクエリの拡張には、WordNet の同義語、上位語、下位語を使うことが有効であると述べている。検索初期段階の洗練されていない単語単位のクエリとは、得られる結果が一様ではないようなクエリである。語彙換言をクエリ拡張に利用している研究として、熊本ら[11]の手法がある。彼らは、クエリを句単位として、その内容語の言い換え候補を Web 検索から入手し、2 つの共起辞書を用いて妥当と判断した言い換え候補で新たなクエリ群を生成している。他にもクエリ拡張に語彙換言を利用している研究は多い。[12,13]

本研究では、単語単位のクエリに対して換言辞書を用いて拡張を行い、元のクエリで獲得した文と拡張したクエリで獲得した文の類似度で評価を行う。

3 用言等換言辞書

用言等換言辞書は、1 章で述べた通り本研究と同様の目的で構築された換言辞書である。形態素解析器 JUMAN (1)の形態素辞書に登録されている品詞が動詞、サ変名詞、形容詞、副詞となっている 12,813 語を換言対象とし、換言対象語を用いた例文を考え、日本語初学者に説明することを想定して平易な語に換言している。換言対象語の意味が分からない、換言語を思いつかない、長い説明が必要な換言になる場合、換言を行わない。また、多義語は換言を行う際に考えた例文が用例として記載されている。換言対象語がサ変名詞である場合は、対象語の後ろに「する」を補い動詞として換言を行っている。格変化が起こる換言の場合や、慣用的表現の一部である場合には、その情報も付加している。

構築された辞書は、12,813 語の換言対象語に対して、10,336 語の換言対が登録されている。

4 普通名詞換言辞書の構築

4.1　作業対象

換言の対象として形態素解析器 JUMAN の形態素辞書に登録されている普通名詞 16,524 語の代表表記を用いた。これにより表記ゆれにより離れた位置にある同じ語を一つにまとめることができ、代表表記が同じ見出し語は同様の換言となる。また、JUMAN の形態素辞書に登録されている普通名詞にはカテゴリが付与されており、そのカテゴリに従った語義についてのみ換言を行った。カテゴリは大分類で、人工物、自然物、場所、組織・団体、人、動物、植物、抽象物、時間、数量、形・模様、色の 12 カテゴリに分かれており、一つの見出し語に対し複数のカテゴリが付与されている場合もある。

4.2　作業手順

換言対象語を見て、その語を作業者の考えで換言する。作業者は日本語を母語とする成人男性である筆者一人である。換言は、日本語初学者からその言葉の意味を問われたときにどのように答えるか、を念頭に置いて換言する。すなわち、作業者の感覚で明らかに単純な語、明らかに難しい語、意味が分からない語は換言しない。

例）明らかに単純な語　：「上」、「男」
　　明らかに難しい語　：「羅」、「稟議」
　　意味が分からない語：「一番鶏」、「健筆」

また、換言する際に内容語を 2～3 語程度に収めるという制限のもとに換言を行う。これは、あまりに長すぎる換言を行わないためである。そのため、元の語の意味を完全に保っているとは限らない。ただし、この制限内で出来る限りの情報を付けて換言する。

例）「折り鶴」→「紙で折った鶴」

4.3　作業基準

4.3.1 換言を行わない場合

以下のような場合に、無記入を許した。

- 換言語が思いつかない　　例)「ストライク」、「アルカリ」
- 元の語の意味が明確でない　例)「村八分」、「氏神」

4.2 節で述べた明らかに単純な語は換言が思いつかない場合、明らかに難しい語は元の語の意味が明確でない場合に含まれる。

これは作業効率を上げるためでもあり、無理な換言を行わないようにするためでもある。

4.3.2 多義語

換言対象語が多義を持っていると判断した場合、語義ごとに換言を行う。ただし、換言対象語の属するカテゴリに従った語義のみについて考える。すなわち、作業者が多義と判断しても付与されているカテゴリにそぐわない意味の場合は換言を行わない。今回は JUMAN の形態素辞書を基に換言を行っているため、付与カテゴリ以外の意味は登録されていないものとして扱うためである。

例)「クラス」　カテゴリ：組織・団体　→「集団」
　　抽象物　　→「階級」
「王冠」　カテゴリ：人工物-衣服　→「王がかぶる飾り」
　　×「瓶のふた」：カテゴリにそぐわない

4.4　作業結果

作業対象の語数と実際に換言を行った項目数、及び無記入とした語数をカテゴリ別に表 1 に示す。4.1 節、4.3.2 節で述べた通り一つの対象語に対して換言結果が一つになるとは限らないため、「換言作成」欄、「無記入」欄の合計と「換言対象」欄の数は一致しない。また、構築した普通名詞換言辞書の一部を本稿の付録に記載している。

JUMAN の形態素辞書に登録されている普通名詞 16,524 語について、約 95%にあたる 16,153 語の換言対を得た。

表 1　換言対象語数と作業結果

カテゴリ	換言対象	換言作成	無記入
人工物	2,610 語	2,557 語	72 語
自然物	453 語	420 語	33 語
場所	1,795 語	1,685 語	111 語
組織・団体	248 語	228 語	20 語
人	1,479 語	1,419 語	66 語
動物	771 語	724 語	47 語
植物	339 語	316 語	23 語
抽象物	6,912 語	6,465 語	435 語
時間	259 語	227 語	33 語
数量	353 語	325 語	29 語
形・模様	135 語	120 語	15 語
色	88 語	84 語	4 語
複数	825 語	1,583 語	92 語
合計	16,267 語	16,153 語	980 語

4.5　考察

無記入については意味が分からない語が三分の二ほどあり、残りは簡単な語にできなかったものであった。意味が分からなかった語には「アフタ」や「建て玉」といった医療や金融などの分野に関する名詞が、簡単な語にできなかった語には「上」や「液体」といった性質、状態を表す名詞や、「シュート」、「キャッチボール」といったスポーツの行為に関する名詞が多く含まれている。専門的な語はその分野特有の語であり、意味を知らなかったり、ほかの言い回しが難しいため換言をしにくい傾向にあると考える。また、性質等を表す語は、説明に用いたりする語であるために簡単な語に換言しにくいと考える。

5 評価実験

5.1　評価方法

今回構築した普通名詞換言辞書の評価として、辞書を用いたクエリ拡張を行う。2章で述べた通り、洗練されていない単語単位のクエリの拡張には同義語、上位語、下位語が有効であることが明らかになっている。今回構築した普通名詞換言辞書と3章で述べた用言等換言辞書を合わせた換言辞書（以下、換言辞書）と日本語 WordNet 同義語データベース Ver.1.0 (2)（以下、WordNet）の両方に見出しとして含まれる普通名詞とサ変名詞の組み合わせをクエリとして、毎日新聞2年分（1999年と2000年）(3)から文の検索を行う。元クエリで獲得した文と換言辞書と WordNet それぞれでクエリ拡張を行って獲得した文で類似度計算を行い、その類似度で評価を行う。類似度が高ければ元クエリと同じような内容の文を獲得していることになるので、クエリ拡張に有効であることがわかる。

5.2　実験方法

5.2.1　元クエリの決定

5.1 節で述べた通り、換言辞書と WordNet の両方で見出し語となっている普通名詞（2,877 語）とサ変名詞（1,021 語）の組み合わせを仮のクエリとし、1999 年と 2000 年の毎日新聞（計 476,586 文）を対象に文検索を行う。

まず対象文を、形態素解析器 JUMAN を用いて形態素に分ける。その後、仮のクエリとした普通名詞とサ変名詞の両方があった場合、その文を照合したクエリに対応する文とする。この際、一文に対し一クエリのみの対応とする。つまり、複数のクエリが同一文に照合する場合でも、初めに照合したクエリに対応した文となる。こうして、文が照合したクエリを元クエリとし、クエリ拡張は元クエリについてのみ行う。ただし、獲得した文が一文のみ、または獲得した文の内容語が全て同じ場合は元クエリから除くこととする。これはのちの類似度計算において文の内容語を用いるので、1パターンでは信頼性に欠けるためである。ここで、内容語とはクエリで用いた語以外の形態素解析で「動詞」、「形容詞」、「普通名詞」、「サ変名詞」、「人名」、「地名」と付与

された単語である。また、「人名」、「地名」と付与された単語については「人名」、「地名」と汎化している。

5.2.2　クエリ拡張による文の獲得

5.2.1 節で決定した元クエリに対して換言辞書と WordNet のそれぞれでクエリ拡張を行い、元クエリの決定に使用した毎日新聞に対し文検索を行う。拡張は、元クエリとなっている普通名詞、サ変名詞と各語に対応した各言語資源の語の組み合わせとなる。また、文検索の際には 5.2.1 節と同様に一文に対し一クエリのみの対応とする。

5.2.3　獲得した文の類似度計算

元クエリで獲得した文とクエリ拡張によって獲得した文で類似度計算を行う。類似度計算には Jaccard 係数と Simpson 係数を用いる。

$$Jacc = \frac{|X \cap Y|}{|X \cup Y|} \qquad (1)$$

$$Simp = \frac{|X \cap Y|}{\min(|X|,|Y|)} \qquad (2)$$

ただし、

X：元クエリで獲得した文の内容語の集合

Y：拡張して獲得した文の内容語の集合

計算は各クエリに対して一文対一文の総当たりで計算する。つまり、元クエリで 5 文、ある拡張で 10 文であれば、50 通りの組み合わせで計算を行うこととなる。計算結果の平均をそのクエリにおけるスコアとし、各クエリでのスコアの平均を言語資源のスコアとする。

5.3　実験結果

5.3.1　元クエリの決定

一文以上の文を獲得した元クエリとして使う普通名詞とサ変名詞の組み合わせの対は 24,510 対、照合した文は 140,604 文であった。下に元クエリの例を示す。
例）「負債　削減」、「肝　手術」、「党　公認」

5.3.2　クエリ拡張による文の獲得

換言辞書の拡張によって追加で獲得した文は 110、237 文、WordNet の拡張によって追加で獲得した文は 110,151 文であり、獲得文数はほとんど変わらなかった。
拡張の例）「負債　削減」
　　換言辞書：「借金　削減」「負債　減らす」「借金　減らす」
　　WordNet：「借入　削減」「負債　カット」「負い目　カット」等

5.3.3　獲得した文の類似度

元クエリで獲得した文集合の類似度で段階的に足切りしたときの、元クエリで獲得した文と各拡張で獲得した文の類似度計算の結果を表 2 に示す。ここで、元クエリでの Jaccard 係数とは、元クエリで獲得した文集合について一文対一文で Jaccard 係数を計算した値の平均である。Jaccard 係数が高いということは得られた結果の内容が一様であることを表しており、この値で足切りを行うことで元クエリでの検索結果が一様であるか否かによる拡張の結果の変化を確認することができる。

表 2　類似度計算結果

	Jacc	Simp	Jacc	Simp	Jacc	Simp	Jacc	Simp	Jacc	Simp
元クエリでの Jacc	≧0.9		≧0.8		≧0.7		≧0.6		≧0.5	
換言辞書	0.0678	0.1544	0.0664	0.1571	0.0687	0.1672	0.0774	0.1799	0.1085	0.2281
WordNet	0.0644	0.1487	0.0673	0.1570	0.0666	0.1584	0.0660	0.1648	0.0981	0.2160
元クエリでの Jacc	≧0.4		≧0.3		≧0.2		≧0.1		≧0.0	
換言辞書	0.1007	0.2340	0.0873	0.2110	0.0908	0.2215	0.0772	0.1934	0.0697	0.1809
WordNet	0.0862	0.1932	0.0834	0.2002	0.0763	0.1906	0.0751	0.1923	0.0713	0.1833

表 3　元クエリの獲得文が 6 文以上の場合の類似度計算結果

	Jacc	Simp	Jacc	Simp	Jacc	Simp	Jacc	Simp	Jacc	Simp
元クエリでの Jacc	≧0.9		≧0.8		≧0.7		≧0.6		≧0.5	
換言辞書			0.0381	0.1005	0.0386	0.1026	0.0991	0.2037	0.1804	0.3402
WordNet							0.0591	0.1923	0.1721	0.3283
元クエリでの Jacc	≧0.4		≧0.3		≧0.2		≧0.1		≧0.0	
換言辞書	0.1492	0.3339	0.1159	0.2721	0.1046	0.2478	0.0799	0.1995	0.0727	0.1896
WordNet	0.1022	0.2110	0.0942	0.2149	0.0783	0.1932	0.0768	0.1966	0.0736	0.1893

元クエリでの Jaccard 係数が 0.6 未満となるクエリを足切りしてからスコアが大きく低下し、元クエリでの Jaccard 係数の値が高いほど各スコアは低くなっている。これは、獲得文数が少ないクエリが多いことによる影響と考えられたため、影響除去のために元クエリでの獲得文数が 1 文増加するごとにどれだけクエリ数が変化するか確認し、5 文以下となる場合を除いて再度類似度計算を行った。その結果を表 3 に示す。5 文以下を除いた場合において、すべてのクエリで類似度計算を行った場合の類似度スコアは、ほとんど変わらない結果となった。しかし、元クエリでの Jaccard 係数で

段階的にクエリを足切りして類似度計算を行った結果を見ると、Jaccard 係数と Simpson 係数のどちらも換言辞書の方が高くなっている。特に元クエリの Jaccard 係数が 0.4 未満を足切りした場合に大きく差がついている。これよりクエリ拡張において、我々が構築した換言辞書は WordNet における同義語と同等以上に有効であることがわかる。

5.4　考察

5.4.1 拡張について

- 普通名詞の拡張

元クエリで 6 文以上獲得した場合において、文を獲得した普通名詞のみを拡張したクエリは 831 個であった。元クエリの普通名詞のカテゴリによって換言辞書による拡張に特徴を見るために、換言辞書によって普通名詞のみ拡張が行われた場合の類似度計算をカテゴリ別で行った。その結果を表 4 に示す。

この結果からわかるようにカテゴリによる特徴は見受けられない。このことから、普通名詞の拡張についてはカテゴリによって用いるか否かの考慮を行わずに利用できると言える。

表 4　普通名詞カテゴリ別類似度

カテゴリ	simp	クエリ数	カテゴリ	simp	クエリ数
全体	0.211	831	人	0.226	225
形・模様	0.191	3	人工物	0.223	28
自然物	0.222	7	-金銭	0.228	12
時間	0.247	5	- 乗り物	0.201	1
場所	0.202	52	-その他	0.221	15
-機能	0.194	7	組織・団体	0.214	69
-施設	0.198	29	抽象物	0.203	435
-自然	0.382	1	動物-部位	0.277	5
-その他	0.200	15	数量	0.183	1

- サ変名詞の拡張

サ変名詞のみの拡張での Simpson 係数は 0.174 となり、普通名詞のみの拡張よりも低くなっていた。しかし、今回使用した換言辞書の一つである用言等換言辞書のサ変名詞に関しては、「サ変名詞＋する」という用言として使われている場合の換言であるため、類似度計算に用いる元クエリで獲得した文のうち「サ変名詞＋する」となっている文とのみ計算する必要がある。元クエリで獲得した文のうち「サ変名詞＋する」となっている文は 91,029 文であった。これを考慮した上で、元クエリで 6 文以上獲得しているクエリに関して類似度計算を行った。その結果 Simpson 係数は 0.181 であった。考慮していなかった場合より、わずかに上昇したが大きな変化は得られなかった。

また、サ変名詞の換言で“評価 → 書く”のような双方向の換言が難しいようなものがあり、このような拡張では類似度スコアが低くなることが考えられたため、筆者の感覚で相互に換言可能か否かに分けて計算を行った。その結果を下に示す。

相互変換　可能　：simp 0.183、クエリ数 522、「逃走」→「逃げる」

不可能：simp 0.175、クエリ数 162、「記録」→「残す」

類似度スコアが低くなる原因の一因ではあるが、このことが原因の核心ではないことがわかる。

普通名詞とサ変名詞の両方を換言して拡張した場合、Simpson 係数は 0.091 という結果となった。拡張してできたクエリを見てみると「債務　返済」が「借金　返す」というように元のクエリより易しい印象になっているものがある。これは、3.2 節で述べたように換言辞書を構築する段階で、日本語初学者からその言葉の意味を問われたときにどのように答えるか、を念頭に置いて換言しているため、クエリの印象が易しくなっている。表現が易しくなることで、内容が同じ記事をこどもニュースから検索することができる可能性があると考えられる。

5.4.2 拡張後の獲得文数について

構築した換言辞書は基本的に見出し語に対して一対一で換言語が登録されている。対して WordNet では見出し語に対して複数の同義語が登録されている。そのため、WordNet の方が拡張してできる語の組み合わせの数は多くなる。実際に元クエリの 24,510 個に対して、換言辞書では 73,530 個、WordNet では 1,074,212 個に拡張されている。この事実だけ見れば、一見 WordNet の方が獲得文数が多くなると考えられ

るが、実際には換言辞書と WordNet の獲得文数はほとんど変わらなかった。クエリがどのように拡張されたかを確認してみると、WordNet による拡張では、「活動」が「写真」というように違和感のある換言になっているものが見受けられた。文章は人が書くものであり、拡張した際に違和感のあるものはあまり文と共起しないはずである。対して、換言辞書はもともと人手で構築したものであるため違和感のあるものはなかった。また、換言先がより人の感覚に近いので、拡張先が少ないにもかかわらず、多くの文と共起していると考える。

5.4.3 類似度スコアについて

前項でもふれた違和感のあるクエリの拡張で獲得した文は、総じて Jaccard 係数と Simpson 係数のどちらも低かった。そのため WordNet でクエリ拡張を行う際は、そのような違和感のある拡張をしないような処理を加える必要がある。換言辞書は人手で換言対を基本的には一対一の対応で構築しているため、余計な処理を行わずに利用できる。

次に、元クエリで獲得した文集合の Jaccard 係数に着目する。元クエリで獲得した文集合の Jaccard 係数が高いと Jaccard 係数と Simpson 係数のどちらも低くなっていた。元クエリで獲得した文集合の Jaccard 係数が高いということは、元クエリで獲得できる文が一様であり、元クエリが洗練されているものであると考えられる。また、元クエリで獲得した文集合の Jaccard 係数が 0.4 前後のクエリはあまり洗練されていないといえる。5.1 節で述べたように今回のような同義語による拡張は、洗練されていない単語単位のクエリに効果があるため、今回得られた結果はその傾向に即していると考えられる。

6. 結論

形態素解析器 JUMAN の形態素辞書を基に人手で普通名詞換言辞書の構築を行った。JUMAN の形態素辞書に登録されている普通名詞約 1 万 7 千語について、約 95%にあたる約 1 万 6 千語の換言対を得た。換言を行わなかった語のうち、三分の二が元の語の意味が分からなかったものであり、残りは簡単な語に換言できなかったものであった。

また、今回構築した普通名詞換言辞書と用言等換言辞書を合わせた換言辞書の評価として文検索を行った。換言辞書と日本語 WordNet 同義語データベースでそれぞれクエリ拡張を行い、元クエリで得られた文と拡張して得られた文で類似度計算を行った。元クエリで得られた文数は 140,604 文、換言辞書による拡張で得られた文数は 110,237 文、日本語 WordNet 同義語データベースによる拡張で得られた文数は 110,151 文であり、換言辞書と日本語 WordNet 同義語データベースで得られる文数は変わらなかった。また類似度計算を行った結果、換言辞書では日本語 WordNet 同義語データベースで拡張を行う場合と同等以上の効果があることがわかった。

今回構築した普通名詞換言辞書と用言等換言辞書を合わせた換言辞書は公開する予定である。

謝辞

本研究を遂行するにあたり、多大なるご指導とご協力を頂きました、長岡技術科学大学の山本和英准教授に深く感謝致します。
また、研究の方針や研究への意見、論文の執筆において協力してくださった山本研究室の皆様に、心より感謝いたします。

使用したツール及び言語資源

(1) 形態素解析器 JUMAN Ver.7.0. 京都大学 大学院情報学研究科 知能情報学専攻

http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

(2) 日本語 WordNet 同義語データベース Ver.1.0 Japanese WordNet Synonyms Database. 独立行政法人情報通信研究機構(NICT)

http://nlpwww.nict.go.jp/wn-ja/

(3) 毎日新聞社. CD-毎日新聞 1999 年度版及び 2000 年度版、 1999　2000.

参考文献

[1] 山本 和英. 換言処理の現状と課題. 言語処理学会 第 7 回年次大会併設ワークショップ(2001.3) 、pp.93-96

[2] 乾 健太郎. 自然言語処理と言い換え. 日本語学、 Vol.26、 No.11、 pp50-59, 2007.11

[3] 山本 和英、 吉倉 孝太郎. 用言等換言辞書を人手で作りました. 言語処理学会第 19 回年次大会 発表論文集 (2013.3)、 pp.276-279

[4] Francis Bond、Timothy Baldwin、Richard Fothergill and Kiyotaka Uchimoto. Japanese SemCor: A Sense-tagged Corpus of Japanese. The 6th International Conference of the Global WordNet Association (GWC-2012).

[5] 柴木 優美、 永田 昌明、 山本 和英. カテゴリ名と記事名の意味属性分類に基づく Wikipedia からの上位下位関係オントロジーの構築. 自然言語処理、Vol.19、No.4、 pp.229-279、 言語処理学会 2012.12

[6] 難波 英嗣、 奥村 学、 新森 昭宏、 谷川 英和、 鈴木 泰山. 特許データベースからのシソーラスの自動構築. 言語処理学会第 13 回年次大会 (2007)、pp.1113-1116.

[7] 安川 美智子、 山田 篤. Web 検索エンジンを用いた用語検索履歴からのシソーラス自動構築手法の評価と改良. DEWS2005 論文集、 5C-i8、 2005.

[8] 梶原智之、 山本和英. 小学生の読解支援に向けた語釈文から語彙的換言を選択する手法. NLP 若手の会 第 8 回シンポジウム、 (発表 23) (2013.9)

[9] 呉 浩東. 語彙・構文的言い換え表現の自動生成. 情報科学研究, (25), pp.95-99. 2008

[10] Ellen M 、 Voorhees. Query Expansion using Lexical-Semantic Relations. In 17th International Conference on Research and development in Information Retrieval (SIGIR’94). p61-69, Springer London, 1994.1.

[11] 熊本 忠彦、田中 克己. 2 種類の共起辞書を用いた語彙的言い換えに基づく Web 検索システム. 人工知能学会論文誌、 Vol.23、 No.5、 pp355-363. 2008

[12] Buscaldi, Davide, Paolo Rosso, and Emilio Sanchis Arnal. A wordnet-based query expansion method for geographical information retrieval. Working notes for the CLEF workshop. 2005.

[13] Hsu, Ming-Hung, Ming-Feng Tsai, and Hsin-Hsi Chen. Query expansion with conceptnet and wordnet: An intrinsic comparison. Information Retrieval Technology. Springer Berlin Heidelberg, pp1-13. 2006.

付録 A 普通名詞換言辞書の一部

今回構築した普通名詞換言辞書の一部を以下に示す。

カテゴリ	代表表記	換言	掲載行数
人工物-その他	合い鍵/あいかぎ	予備の鍵	9
人工物-その他	合い口/あいくち	鍔がない短刀	18
人工物-乗り物	愛車/あいしゃ	自分の車	30
人工物-食べ物	アイス/あいす	氷菓子	39
人工物-その他	アイテム/あいてむ	道具	63
人工物-衣類	合い服/あいふく	制服	76
人工物-その他	アイロン/あいろん	しわを伸ばす家電	85
人工物-食べ物	赤米/あかごめ	赤いお米	133
人工物-その他	空き缶/あきかん	空の缶	176
自然物	亜鉛/あえん	金属	98
自然物	赤錆び/あかさび	赤い錆	134
自然物	赤土/あかつち	赤い土	143
自然物	灰汁/あく	灰を入れた水	201
自然物	朝露/あさつゆ	露	292
自然物	朝日/あさひ	朝上ってくる太陽	297
自然物	アスベスト/あすべすと	石綿	359
自然物	汗/あせ	体液	368
自然物	汗水/あせみず	汗	372
場所-自然	アイランド/あいらんど	島	84
場所-その他	アウトドア/あうとどあ	屋外	89
場所-施設部位	明かり取り/あかりとり	天窓	160
場所-その他	亜寒帯/あかんたい	寒い地域	163
場所-施設部位	上り口/あがりぐち	入口	168
場所-その他	空き地/あきち	あいた土地	183
場所-施設	空き家/あきや	空いている家	190
場所-その他	朝市/あさいち	朝に開催される市	280
場所-自然	浅瀬/あさせ	浅い水たまり	289
組織・団体	委/い	×	877
組織・団体	医学部/いがくぶ	医療を学ぶ学部	973
組織・団体	遺族/いぞく	故人の家族	1204
組織・団体	一群/いちぐん	一つの群れ	1265
組織・団体	一族/いちぞく	一つの部族	1290
組織・団体	一団/いちだん	団体	1293
組織・団体	一味/いちみ	組織	1321
組織・団体	一門/いちもん	門下全員	1329
組織・団体	一家/いっか	家族	1351
組織・団体	一行/いっこう	グループ	1373
人	相方/あいかた	コンビの相手	10
人	愛妻/あいさい	妻	28
人	愛児/あいじ	子供	35
人	愛人/あいじん	浮気相手	37
人	あいつ/あいつ	相手	56
人	相手/あいて	対する人	60
人	アイドル/あいどる	もてはやされる人	70
人	相棒/あいぼう	コンビの相手	77
人	赤子/あかご	赤ちゃん	132
人	赤ちゃん/あかちゃん	幼児	141
動物	愛犬/あいけん	自分の犬	20
動物	青大将/あおだいしょう	蛇	115
動物	青虫/あおむし	芋虫	123
動物-部位	垢/あか	汚れ	126
動物-部位	赤毛/あかげ	赤い毛	131
動物-部位	赤身/あかみ	赤い身	152
動物	赤虫/あかむし	赤い虫	153
動物-部位	赤ら顔/あからがお	赤い顔	154
動物	揚げ羽蝶/あげはちょう	蝶	270
植物	藍/あい	青い花	4
植物	葵/あおい	花	102
植物	青黴/あおかび	青いカビ	105

植物	青菜/あおな	葉野菜	117	数量	緯度/いど	南北の位置の単位	1506
植物-部位	青葉/あおば	緑の葉	120	数量	インチ/いんち	長さの単位	1805
植物	アカシア/あかしあ	木	136	形・模様	市松/いちまつ	チェック柄	1320
植物	茜/あかね	赤い花	147	形・模様	一環/いっかん	一つの輪っか	1358
植物	赤松/あかまつ	針葉樹	151	形・模様	一線/いっせん	一本の線	1413
植物	秋草/あきくさ	秋の草花	177	形・模様	一直線/いっちょくせん	一本の直線	1429
植物	麻/あさ	布材料の草	278	形・模様	後ろ姿/うしろすがた	後ろからみた姿	1956
抽象物	愛/あい	感情	3	形・模様	絵柄/えがら	絵	2428
抽象物	哀感/あいかん	悲しい感情	12	形・模様	円形/えんけい	円	2580
抽象物	哀歓/あいかん	悲しみと喜び	13	形・模様	円周/えんしゅう	円の周り	2594
抽象物	合気道/あいきどう	武道	16	形・模様	円陣/えんじん	円形の陣	2611
抽象物	愛嬌/あいきょう	かわいらしい態度	17	色	藍色/あいいろ	濃い青	7
抽象物	愛国/あいこく	国を愛する気持ち	24	色	青/あお	空の色	99
抽象物	アイコン/あいこん	象徴	26	色	青色/あおいろ	青	104
抽象物	哀愁/あいしゅう	悲しい雰囲気	31	色	赤/あか	リンゴの色	127
時間	合間/あいま	間の時間	78	色	赤色/あかいろ	赤	130
時間	暁/あかつき	明け方	142	色	茜色/あかねいろ	夕焼けの色	148
時間	曙/あけぼの	明け方	260	色	浅緑/あさみどり	淡い緑	299
時間	当たり年/あたりどし	都合のいい年	412	色	飴色/あめいろ	茶色	642
時間	後先/あとさき	これから先	494	色	暗紅/あんこう	暗い赤	804
時間	安息日/あんそくび	休日	828	色	イエロー/いえろー	黄色	937
時間	十六夜/いざよい	夏の夜	1114	人	相手方/あいてかた	相手	61
時間	一期/いちご	一生	1270	場所-機能	相手方/あいてかた	相手の側	61
数量	価/あたい	価格	385	植物	青海苔/あおのり	海藻	119
数量	値/あたい	数値	386	人工物-食べ物	青海苔/あおのり	細かくした海苔	119
数量	あまた/あまた	たくさん	607	人工物-その他	証/あかし	証拠	135
数量	アール/あーる	×	871	抽象物	証/あかし	証明	135
数量	言い値/いいね	宣言した値段	909	組織・団体	アカデミー/あかでみー	教育機関	144,145
数量	一元/いちげん	×	1268	場所-施設	アカデミー/あかでみー	学校	144, 145