普通名詞換言辞書の構築

山形 祐輝　　山本 和英
長岡技術科学大学　電気系
{yamagata、 yamamoto}@jnlp.org

1 はじめに

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。また、言語処理の分野において、コーパスや辞書を用いた研究が盛んであり、数多くの言語資源が作成されてきた。

換言辞書に類似した言語資源として日本語WordNet [2]、柴木ら[3]が提案しているような is-a 関係オントロジーによるシソーラス、国語辞典の語釈文等が挙げられる。実際に、語彙換言の典型的手法として、これらを換言知識として用いるものがある。梶原ら[4]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。しかし、このような言語資源を用いた手法の場合、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。また、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言語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[5]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、本稿では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 作業内容

2.1 作業対象

換言の対象として形態素解析器 JUMAN (1)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。代表表記が同じ見出し語は同様の換言となる。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 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。

2.2 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。作業者は著者のうちの一人である。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。

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。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」

2.3 作業基準

2.3.1　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」
- 元の語の意味が明確でない　例)「羅」

2.2 節で述べた明らかに単純な語、明らかに難しい語は換言が思いつか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。

2.3.2　多義語

換言対象語が多義であると判断した場合、換言対象語を複製して、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、抽象物
　　→「集団」、「階級」

3 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。2.1 節、2.3.2 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。

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。

4 評価実験

4.1 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

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

書を用いたクエリ拡張を行う。Ellen M ら[6]は情報検索において、得られる結果が一様ではない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、下位語が有効であると述べている。換言辞書は同義表現をまとめたものであり、WordNet の同義語と同様にクエリ拡張に有用なはずである。

そこで、今回構築した普通名詞換言辞書と 1 節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し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、元クエリでは獲得できなかった文を多く入手可能になり、クエリ拡張に有効であることがわかる。

4.2 実験方法

4.2.1　元クエリの決定

4.1 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。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。また、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。

4.2.2　クエリ拡張による文の獲得

4.2.1 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 4.2.1 節と同様に一文に対し一クエリのみの対応とする。

4.2.3　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。

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

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

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、拡張して 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。その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。

4.3 実験結果

4.3.1　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。

例）「学校　存在」、「肝　手術」、「党　公認」

4.3.2　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。

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

拡張の例)「学校　存在」
　換言辞書　「学校　いる」
　WordNet　「学校　いる」「学校　ある」
　　「学院　いる」「学院　ある」
　　「学園　いる」「学園　ある」

4.3.3　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。これは、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ためである。

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。これより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 考察

5.1 普通名詞換言辞書について

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。

5.2 評価実験結果について

- 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。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、実際に WordNet の方が多くの語に拡張されているが、「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。

- 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。4.1 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 おわりに

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。

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。その結果、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。

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] 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).

[3] 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

[4] 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)

[5] 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

[6] 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.

使用した言語資源及びツール

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.