観点絞り込み検索法による類義語生成

金杉 友子† 笠原 要†† 松澤 和光††

†NTT アドバンステクノロジ(株) ††NTT コミュニケーション科学研究所

1 はじめに

知識が不完全でも概略的な答を導く「アバウト推論」[5]の研究の一環として、単語の意味を体系的に知識ベース化した「概念ベース」の構築[3]を行なっている。

概念ベースでは、N 個の語の各々の意味(概念 g)を、同じく N 個の語から選んだ g の特徴を表す語(属性 p)、および p の重要度 q の対の集合で表す。すなわち

$$概念g = \{(p1,q1),(p2,q2)\ldots(pN,qN)\} \quad (1)$$

ここで各 p,q は、国語辞典等における見出し g に対する語義文から、自立語とその出現頻度として機械的に抽出することができ、実際に4万語規模の概念ベースを構築済みである。

概念ベースでは、各語は N 次元ベクトルとして表現されているから、これらの比較によって単語間の類似度を計算できる。ただし実際には、N=4万語に比べ、抽出された属性の数が少な過ぎるため、適当なシソーラス(ここではALT シソーラス[1])を用いて属性を M 個のカテゴリに圧縮し、M 次元空間でのベクトル同士の成す角の余弦により類似度を表す。

また、単語間の類似度はその単語が用いられている状況や文脈(観点)に応じて変化する。例えば、概念「馬」は「動物」を観点とした場合は「豚」と似ているが、「乗り物」の観点では「自動車」に似ている。そこで観点として指定した概念に応じて、単語間の類似度を計算する「観点変調方式」[2]を開発した。これにより、ある観点においてある単語と類似した単語を、概念ベースに含まれる全単語中から検索することが出来る。

しかし、実際に検索を行なうと検索結果には「観点によらず常に概念に類似している語」(類義語)が多く出てしまう(表 1)。また、類似度自体に絶対的な意味がないため、どこまでを類似語とすべきかも不明である。そこで、観点として指定された単語が「類似語検索の際重視すべき属性」(観点属性)であると考え、観点属性を含む語のみを類似語として検索する「観点絞り込み検索」方式[4]を開発した。

ところが、この方式にもいくつかの問題点がある。まず、観点属性そのものを持たないと類似語と見做さないため、検索結果に抜け落ちが多い。また、概念の類義語はあまり出ないが、観点に関連する語が多く出過ぎる。例えば、観点と概念が同じカテゴリ分類の場合(表 2.1)、ある程度両者の類似語は重なっている筈だが、それでも概念の類似語としては不適切な観点に関連する語(「乗る」、「乗り手」等)が結果に出てしまう。一方、観点と概念が異なるカテゴリ分類の場合は(表 2.2)、概念と関係ない観点の類義語(「深海」、「青海原」等)が多く検索される。

本報告では、以上の問題を解決するため「観点絞り込み検索」方式自体を利用して「類義語」を生成し、これによって類似語検索方式を改善する手法を提案する。

表1：通常の類似語検索結果

観点「乗り物」／概念「馬」

順位	検索結果	類似度
1	馬	1.0000
2	名馬	0.9083
3	愛馬	0.9077
4	馬糞	0.9050
5	馬蹄	0.9043
6	嘶く	0.8921
7	裸馬	0.8857
8	馬蹄形	0.8744
9	じゃじゃ馬	0.8715
10	調教	0.8593

表2：観点絞り込み検索結果

(表2.1)

観点「乗り物」／概念「馬」

順位	検索結果	類似度
1	車馬	0.5150
2	乗る	0.3907
3	乗り手	0.3872
4	乗り合い	0.3697
5	曲乗り	0.3520
6	相乗り	0.3266
7	乗り換え	0.3124
8	乗り掛かる	0.2987
9	牛車	0.2951
10	同乗	0.2920

(表2.2)

観点「海」／概念「鯨」

順位	検索結果	類似度
1	鯨	1.0000
2	海豚	0.7745
3	千尋	0.7321
4	深海	0.7198
5	鱸	0.6859
6	一衣帯水	0.6520
7	鰆	0.6491
8	青海原	0.6462
9	鰊	0.6418
10	茫洋	0.6393

2 類義語生成方式

2.1 定義

方式提案に先立ち、用語を定義する。

[1] 観点絞り込み検索:

$$K(a,b) = \{b_1, b_2, \ldots\} \qquad (2)$$

b_n は a を観点とする b の類似概念として、観点絞り込み検索で検索された結果。

[2] 観点概念同語絞り込み検索:

$$W(a) = K(a,a) \qquad (3)$$

[1] の観点絞り込み検索を観点と概念を同じ a にして行なった場合の結果。

[3] 多重検索:

$$T(A,b) = \bigcup_A K(a,b) \mid a \in A \qquad (4)$$

集合 A 中の要素各々を観点として [1] を行ない結果をマージしたもの。

[4] 再帰検索:

$$S(a) = T(W(a), a) \qquad (5)$$

[5] 下位カテゴリ集合 $X(a)$:

a のシソーラスにおけるカテゴリおよびその下位カテゴリの集合。

[6] 多重カテゴリ集合 $C(A)$:

集合 A 中の要素各々のカテゴリをマージした集合。

[7] カテゴリ絞り:

$$D(A,C) = \{a \mid a \in A, c(a) \in C\} \qquad (6)$$

A のカテゴリ $C(A)$ がカテゴリ集合 C に含まれるものだけを抽出。

[8] 相互属性絞り:

$$E(A,b) = \{a \mid a \in A, b \in Z(a)\} \qquad (7)$$

a の属性集合 $Z(a)$ 中 b を属性に含むものだけを抽出。

2.2 方式の基本的な考え方

ここでは類義語を「概念本来の観点で類似している語」と考え、観点と概念を同じ語にして類似語検索を行ない類義語を生成する。類似検索方式としては、検索結果が多く出過ぎないように観点絞り込み検索を採用する。しかし、この方式では 1 章で述べたように、不適切な語が検索されたり、必要な語が抜け落ちたりする。そこで、不適切な語は元の語のカテゴリ分類等を利用して削除し、抜け落ちは検索を再帰的に行なって補充することにする。以上から、類義語生成は以下の 4 ステップで構成される。

step1) 観点概念同語検索を行なう。

step2) カテゴリ分類等を利用して、検索結果を絞り込む。

step3) 再度類似検索を行ない、必要な語の抜け落ちを補充する。

step4) 再度カテゴリ分類を利用して、検索結果を絞り込む。

ただし、概念のカテゴリ分類が具体物・抽象物(以下「具体」)か、抽象的な事・抽象的関係(以下「抽象」)かでカテゴリへの含まれ方が違うため、step2) と step4) については各々 2 通りの手法に分けた。

step2) では、

- 概念が具体の場合

  シソーラスの分類が類義語生成に適切で、1 つのカテゴリの上位下位の中に類義語が収まる。よって、下位カテゴリ集合による絞り込みを行なう。

- 概念が抽象の場合

  シソーラスの分類が類義語生成には不適切。1 つのカテゴリの上位下位の中に類義語が収まらない。よってカテゴリ集合による絞り込みは行なわず、相互属性検索により絞り込む。相互属性検索の前段階で用いる再帰検索は 1 つのカテゴリ分類に関係なく類似語を検索するため、類義語が複数のカテゴリ系統にわたって分類される抽象語においては適切である。

step4) では、

- 概念が具体の場合

  step2) 同様下位カテゴリ集合による絞り込みを行なう。

- 概念が抽象の場合

  step2) により絞り込まれた類義語候補は複数のカテゴリに分類されている。そこで、概念と step2) の類義語候補の多重カテゴリ集合による絞り込みを行なう。抽象のカテゴリ分類は粗く、複数のカテゴリ系統を扱っている。下位カテゴリまで考えると該当語が多くなり絞り込みの効果が薄くなるため、あえて見ないものとした。

2.3 類義語生成方式

以上に従って、実際の方式を示す。(図 1 参照)概念 a が具体の場合の類義語集合 P は、

$$\text{step1-a)} \quad A = W(a)$$
$$\text{step2-a)} \quad A' = D(A, X(a))$$
$$\text{step3-a)} \quad B' = T(A', a)$$
$$W' = A \cup B'$$
$$\text{step4-a)} \quad P = D(W', X(a))$$

となる。
また、概念 a が抽象の場合の類義語集合 Q は、

$$\text{step1-b)} \quad A = W(a)$$
$$\text{step2-b)} \quad A'' = E(S(a), a)$$
$$\text{step3-b)} \quad B'' = T(A'', a)$$
$$W'' = A \cup B''$$
$$\text{step4-b)} \quad Q = D(W'', Y(a))$$
$$Y(a) = C(a) \cup C(A'')$$

となる。

	a : 具体	a : 抽象
step1	観点概念同語カテゴリ絞り検索 類似語群A	
step2	下位カテゴリ絞り 類似語群A'	相互属性絞り 類似語群A"
step3	多重検索 類似語群W'	類似語群W"
step4	下位カテゴリ絞り 類義語群P	多重カテゴリ絞り 類義語群Q

図 1: 類義語生成過程

2.4 方式の適用例

提案方式を適用した例を表 3に記す。
これらの結果から、提案方式による類義語生成の効果が伺える。

3 類似語検索への適用

3.1 方式の基本的な考え方

前章の類義語を実際の類似語検索に適用する方式を考える。以下の 4 ステップで構成される。

step1) 観点 a、概念 b で観点絞り込み検索を行なう。

step2) 上記検索での抜け落ちを補充するため、類義語生成の step2) で生成した a の類似語群 (A'/A'') を観点に概念 b で観点絞り込み検索を行なう。

step3) 上記 step1) と step2) をマージ。

step4) 上記 step3) の結果に a の類義語群 (P/Q) を作用。

ここで step2) は、2 章で述べたように、具体 / 抽象で A'/A'' が異なる。
また、step4) は、1 章で述べたように、観点と概念が同じカテゴリ分類 / 違うカテゴリ分類で作用のさせ方が異なる。

表 3 : 類義語生成結果

「乗り物」の類義語

順位	検索結果	類似度
1	車	2.1356
2	乗り物	1.6986
3	輿	1.5225
4	車輌	1.4352
5	新車	1.3832
6	手車	1.3584
7	始発	1.3439
8	牛車	1.2806
9	快速	1.2599
10	車体	1.2543

「海」の類義語

順位	検索結果	類似度
1	内海	5.0933
2	海	3.6247
3	潟	2.3889
4	外海	2.2907
5	外洋	2.2269
6	青海原	2.0176
7	海溝	1.9793
8	大洋	1.8768
9	大海	1.8502
10	洋上	1.8128

「赤い」の類義語

順位	検索結果	類似度
1	赤い	4.4044
2	赤	4.1945
3	朱	3.2865
4	丹	2.5758
5	紅白	2.4604
6	紅	2.2942
7	丹青	1.7457
8	葡萄色	1.7418
9	暖色	1.7413
10	緋	1.7306

「感情」の類義語

順位	検索結果	類似度
1	感情	5.6398
2	思う	2.7956
3	気色	2.5285
4	気分	2.2876
5	心地	2.2570
6	情緒	2.2452
7	憂い	1.8138
8	気	1.8071
9	心気	1.7425
10	気持ち	1.6705

表 4 ：類似語検索への適用結果

観点「乗り物」/概念「馬」

順位	検索結果	類似度
1	車馬	0.9273
2	牛車	0.8825
3	車	0.8678
4	快速	0.8272
5	手車	0.7111
6	車上	0.6630
7	輿	0.5913
8	三輪車	0.5183
9	新車	0.4807
10	馬車	0.4214

観点「海」/概念「鯨」

順位	検索結果	類似度
1	鮪	1.9591
2	鯨	1.8125
3	鮫鰊	1.6893
4	鰈	1.5901
5	鱸	0.6859
6	海図	1.4362
7	水位	1.4042
8	鴎	1.3376
9	鱈	1.3143
10	高潮	1.2139

観点「感情」/概念「悲しい」

順位	検索結果	類似度
1	思う	2.7654
2	愛い	2.1353
3	哀れ	1.7851
4	暗然	1.7839
5	胸	1.7391
6	涼しい	1.6691
7	寂しい	1.6620
8	悲しい	1.5219
9	情緒	1.4834
10	気	1.4694

観点「赤い」/概念「林檎」

順位	検索結果	類似度
1	赤札	2.3036
2	紅葉	2.2043
3	赤外線	1.8376
4	鈴蘭	1.6581
5	唐辛子	1.6378
6	千両	1.6170
7	匂い	1.6133
8	椿	1.6129
9	夕焼け	1.5616
10	赤身	1.5445

- 同じカテゴリ分類の場合

  観点と概念が同じカテゴリということは、両者の類似語群はそのカテゴリ分類に収まる語においては重なる。よって、カテゴリによって絞り込みをかけた観点の類義語群を step3) の結果から抽出する。

- 違うカテゴリ分類の場合

  従来の観点絞り込み検索では「観点の類義語」が上位に検索されてしまうことが問題であった。そこで、観点の類義語群を step3) の結果から削除する。

3.2 類似語検索方式

実際の検索方式は以下の通り。a が具体 / 抽象か、b が a と同じカテゴリ / 違うカテゴリかで 4 通りに分かれる。

step1) $M = K(a, b)$

step2) a が具体の場合: $M' = T(A', b)$
a が抽象の場合: $M'' = T(A'', b)$

step3) a が具体の場合: $Z' = M \cup M'$
a が抽象の場合: $Z'' = M \cup M''$

step4) (図 2 参照)

なお、A', A'' は類義語生成の過程で算出した a の類似語群。P は a が具体の場合の a の類義語群。Q は a が抽象の場合の a の類義語群。

3.3 適用例

実際に適用した結果は表 4の通り。

これらの結果から、観点が概念と異なるカテゴリの場合の検索にはまだ改良の余地があるものの、類義語の類似語検索への応用の有効性が伺える。

観点の分類 / 観点概念	具体	抽象
同じカテゴリ	(1) Z' ∩ P	(3) Z" ∩ Q
違うカテゴリ	(2) Z' − P	(4) Z" − Q

図 2: 類義語検索への類義語適用方式

4 終わりに

本稿では、観点絞り込み検索を用いた類義語の生成と、その類義語を類似語検索に応用する方式を示し、両方式ともに実際適用して、その有効性の見込みを得た。使用したシソーラス以外への適用の可能性、また観点が概念と異なるカテゴリの場合の類似語検索手法の改良など、今後の課題としたい。

参考文献

[1] 池原, 宮崎, 横尾: 日英機械翻訳のための意味解析辞書, 情処研報, *NL84-13, pp. 95-102* (1991)

[2] 笠原, 松澤, 石川, 河岡: 観点に基づく概念間の類似性判別, 情処論, *Vol. 35, No. 3, pp. 505-509* (1994)

[3] 笠原, 藤本, 松澤, 石川: 精錬化に基づく概念ベース構成法, 信学技報, *DE95-7, pp. 49-56* (1995)

[4] 笠原, 松澤: 概念ベースを用いた常識語の類似検索, 信学技報, *AI95-25, pp. 23-29* (1995)

[5] 松澤, 石川, 湯川, 河岡: アバウト推論 — 常識的な推論を目指して —, 人工知能学会研究会資料, *SIG-FAI-9401-1, pp. 1-8* (1994)