3G－9

シソーラスの多次元化のための観点の半自動抽出法

片桐康裕　宮崎正弘

新潟大学大学院工学研究科

1　はじめに

従来のシソーラスは、単語に上位下位の関係づけを行なうことによって、平面的なツリー構造で構成されている。また、分類観点として「ｉｓａ」（部分的に「ｈａｓａ」）が用いられている。しかし、このような単一観点による分類では、意味解析などの処理を行なう場合必ずしも満足ゆく結果が得られない。このような問題を解決するためには、種々の観点によって単語を分類した多次元のシソーラス [1] が必要となる。多次元のシソーラスを作成する際に重要となるのが語を分類する観点の抽出である。本稿では、日本語の特徴である漢字に着目した、多次元シソーラス作成のための観点の半自動抽出法について述べる。

2　名詞性二字熟語の構成

漢字は表意文字であり、それぞれ一文字が基本概念を持っている。また、名詞の多くが二字熟語で構成されており、ほとんどが名詞＋名詞（金山、雨雲）、形容詞＋名詞（大雪、強火）、動詞＋名詞（暴風、荒海）に分類される。このような名詞性二字熟語は、前方の漢字が後方の名詞性一字漢字（主名詞）を連体修飾する関係をとるものが大部分を占める。この場合、前方の漢字が主名詞の観点を表す。例えば、『風』と『春風』の関係を考えると『風』の方が広い概念を持つ。つまり、『風』に何らかの制約を加えることにより『春風』となる。この制約が前方一字漢字の概念であり、主名詞の観点である（図1参照）。

しかし、名詞＋名詞の名詞性二字熟語の場合、連体修飾関係以外に互いに類語の場合（河川、岩石）や互いに対語の場合（上下、左右）がある。このような場合には観点は存在しない。

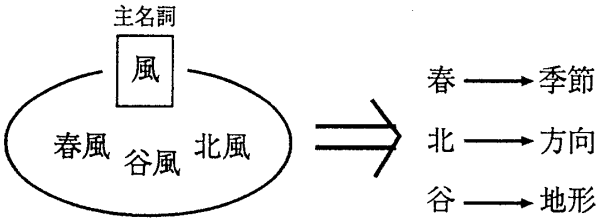

図 1: 二字熟語における観点抽出

3　観点抽出法

3.1　二字熟語の獲得

主名詞を含む二字熟語の獲得を行なう。データベースとして角川類語新辞典 [2] を用いる。

まず、主名詞が角川類語新辞典のどこにあるかを検索する。それには、漢字シソーラス [3] を利用する。漢字シソーラスには一字漢字が角川類語新辞典や既存シソーラス [4] のどこに存在するかという情報や、一字漢字の品詞情報（名詞性、形容詞性、動詞性のどれであるか）、また、多品詞である場合にはその一字漢字における品詞の優先順位などの情報が収録されている。

次に主名詞が存在するノードから主名詞を含む二字熟語をすべて獲得する。多義の発生から一字漢字が角川類語辞典では多ノードに分布していることが多い。この場合は、別のものとして考えて抽出する。

3.2　品詞分類

獲得した主名詞を含む二字熟語から、主名詞を修飾する前方の一字漢字を抜き出す。この時、主

A Semi-Automatic Extracting Facets for Multi-Demensional Thesaurus

Yasuhiro Katagiri, Masahiro Miyazaki
Niigata University

名詞と類語や対語の漢字は含まない。これらの一字漢字を漢字シソーラスを用いて品詞分類する。漢字シソーラスにある品詞優先順位は、一字漢字が単独に出てきた場合の順位なので、すべての品詞について考える。

本稿では二字熟語が名詞＋名詞の場合について考える。つまり、前方が名詞性一字漢字の場合における観点の抽出を行なう。

3.3 クラスタリングによる観点抽出

獲得した名詞性一字漢字に対し既存のシソーラスを用いてクラスタリングを行なう。漢字シソーラスを用い、一字漢字がシソーラス上のどのノードに分類されているか検索する。この時も多義の発生により、一字漢字が多ノードに分類されていることが多い。この場合、最初は漢字シソーラスにおける優先順位が1位のものにおいてクラスタリングを行なう。また、角川類語新辞典は3桁のコードにアルファベットを付与することで細分化されている。クラスタリングの際はアルファベット別で行なう。

クラスタリングの方法は、同じノード内にある漢字をまとめる。そしてそのノードの下位に漢字があるか検索し、ある場合にはまとめておく。次に上位に一つ上がり、これを繰り返す。

上位に上げ過ぎてしまうと抽象度が高すぎて、主名詞の観点としては適当でない。そこでストッパーを付けることにする。ストッパーを付ける位置は、具体物・抽象物、自然物・人工物、生物・無生物の下とする。図2に主名詞を『風』とした時の観点抽出例を示す。観点名は、シソーラスのノード名とする。

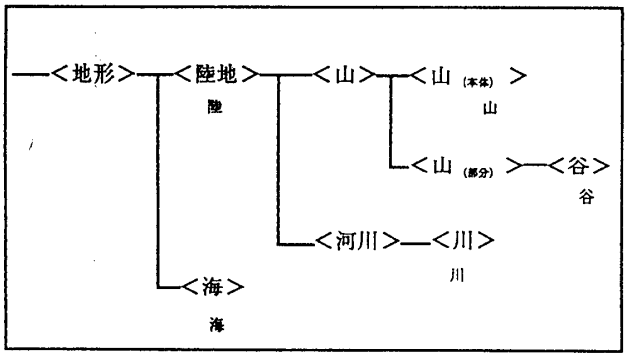

図 2: 二字熟語における観点抽出

この方法を用いることにより、『風』の観点として［地形］［暦日］［場所］等が得られる。

3.4 観点の抽出失敗例

次のような場合では、観点の自動抽出を行なうと関係ないものが抽出される場合がある。

- 種を表す二字熟語を獲得した場合
  白鳥　雷鳥　文鳥
- 偶然重なってしまう場合
  台風　和風　強風
- 明らかに観点とはならないもの
  銀行　砂利　台風

主名詞を『風』とした場合には「台」「和」「強」が＜数量＞でまとまる。しかし、「台」は常用漢字表による漢字制限の実施にともなう書き換え語である（颱風→台風）。また、「和」「強」は形容詞としてまとめられる語である。

このような、観点とならない漢字は計算機では判別できないので、人手で省くことになる。

4 おわりに

二字漢字熟語に着目し、主名詞における観点の抽出法について述べた。今後は、形容詞・動詞のクラスタリングを行なった後に、多次元的なシソーラスの半自動生成へ活用する予定である。

謝辞

名詞意味属性体系データ（名詞シソーラス）を提供して下さったＮＴＴコミュニケーション科学研究所の池原悟氏に深謝する。

参考文献

[1] 川村、宮崎：語を種々の観点から分類した多次元シソーラス、第４８回情報処理学会全国大会、No.3Q-2(1994)

[2] 大野、浜西：角川類語新辞典、角川書店、(1981)

[3] 川村、宮崎:既存シソーラスを利用した漢字シソーラスの半自動生成法、信学技報、NLC93-59、pp37-44(1993)

[4] 池原、宮崎、横尾：日英機械翻訳のための意味解析用知識とその分解能、情報処理学会論文誌、Vol.34、No.8、pp1692-1704、(1993)