既存のシソーラスを利用した仮想シソーラスの構築

小林将　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

従来のシソーラスは、そのほとんどが上位 - 下位関係のような単一の分類観点に基づき一次元的な木構造で構築されているが、単語や概念の分類観点には上位 - 下位関係以外に部分 - 全体関係、反意関係、静的属性、動的属性、機能、源泉など多くの観点がある。これらの情報を表現するためには、単語を複数の観点から分類した多次元シソーラス [1][2] が必要である。しかし、多次元シソーラスの構築には多大な労力が必要となることが予想されるため、可能な限り人手を用いずに自動構築することが望ましい。

上位－下位関係のシソーラスであっても、木構造や各ノードの名称 (分類観点) を手掛かりにある程度多次元的な見方をすることで、多次元シソーラス自動構築の基となることが期待できる。そのためには、豊富な語彙を持ち、末端ノードの粒度が細かく、かつ適切な分類観点の明示されたシソーラスが必要となる。

本稿では、多次元シソーラスの構築を目指して、様々な特徴を持った既存の上位 - 下位構造のシソーラスを複数組み合わせ、仮想的に木構造・語彙・分類粒度・分類観点共に優れたシソーラスを構築するための方法を提案する。

2　今回使用するシソーラス

今回使用する既存のシソーラスは以下の通り。

- 日本語語彙大系 [3]
- 角川類語新辞典 [4]
- 分類語彙表 [5]

以下、これらのシソーラスの特徴について述べる。

2.1　日本語語彙大系

自然言語処理用に作られたシソーラス。上位－下位関係の非常に深い木構造を持ち、語彙も豊富だが、末端ノードの粒度が粗い。

以下、「岩類」と省略する。

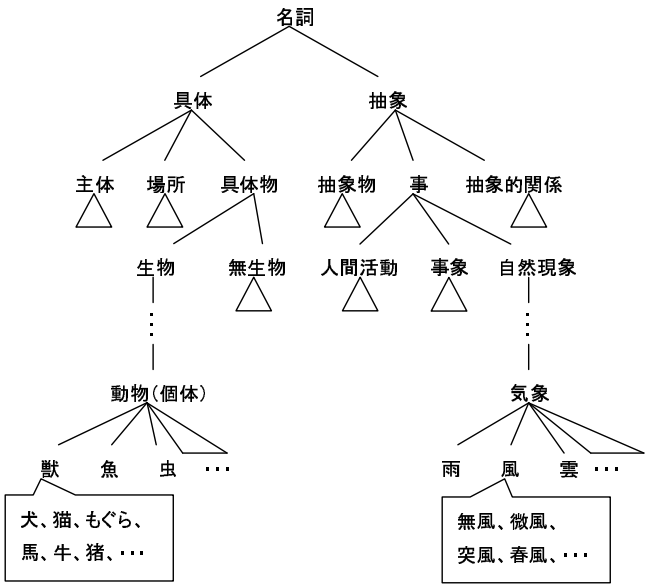

図 1：日本語語彙大系

2.2　角川類語新辞典

人間用のシソーラス。木構造は岩類に比べて浅いが、末端の粒度は比較的細かく、それぞれに分類観点が明示されている。

以下、「角類」と省略する。

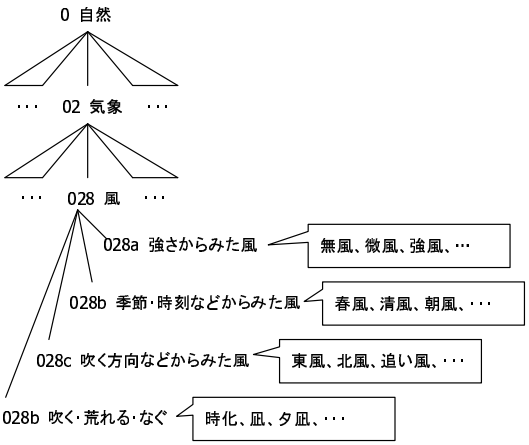

図 2： 角川類語新辞典

2.3 分類語彙表

人間用のシソーラス。上位－下位関係の木構造を持つが、非常に浅く、横に広い。末端ノードの粒度は非常に細かいのだが、分類観点が明示されていない。

以下、「国類」と省略する。

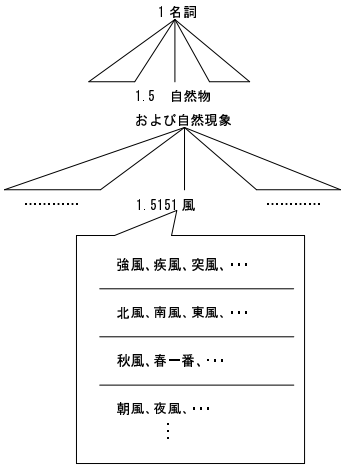

図 3： 分類語彙表

3 仮想シソーラスの構築

岩類は自然言語処理に使う場合、木構造としては優れているが、末端ノードの粒度が粗い。対して、角類・国類（以下、「他シソーラス」）は木構造は岩類に劣るものの、末端ノードの粒度が細かい。そのため、岩類の末端ノードを他シソーラスを利用して細分化することにより、木構造・末端粒度共に優れたシソーラスが構築できると考えられる。

具体的には、岩類の末端ノードの各語を他シソーラスから検索し (図 4)、一致した語を含む他シソーラスのノードを岩類の末端ノードの下に接ぎ木する。ここで、一致した語を含むノードを全て追加したのでは、上位／下位関係として適切でないノード（図の例では“紛争”等）が接ぎ木されることがある。そこで、他シソーラスで一致した語数、およびノード内の全語数と一致語数との割合により制限をかける。一致語数の割合が一定以上のノードのみを接ぎ木することで、図 5 のように適切に細分化された仮想シソーラスが得られる。

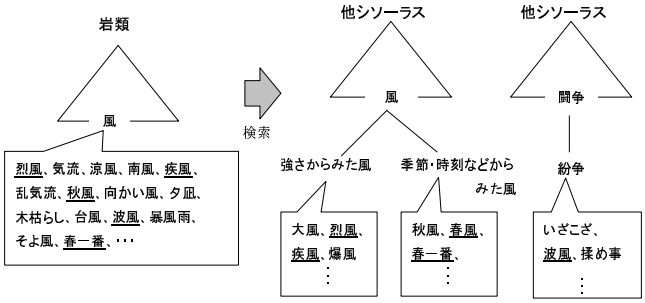

図 4： 仮想シソーラスの構築過程

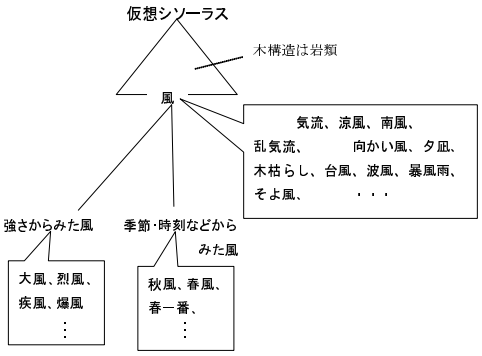

図 5： 仮想シソーラス

3.1 国類の末端ノードの分類観点の推測

国類の末端ノードは非常に粒度が細かいが、ノード名（分類観点）が明示されていない。そこで、比較的粒度の細かい角類を用いて、国類の末端ノードに仮のノード名を付与する。

具体的な方法としては、国類のある末端ノード内の語それぞれについて、角類から検索する。その結果、最も多くの語が一致した角類のノード名を国類の末端ノードの仮のノード名として採用する。

例を図6に示す。国類“風”の末端ノード内の語が他シソーラスの“強さから見た風”内の語と最も多く一致した場合、国類“風”の末端ノードに“強さから見た風”というノード名を付与する。

なお、他シソーラスに一致する語が存在しない、もしくは著しく少なかった場合は、ノード名なしとする。

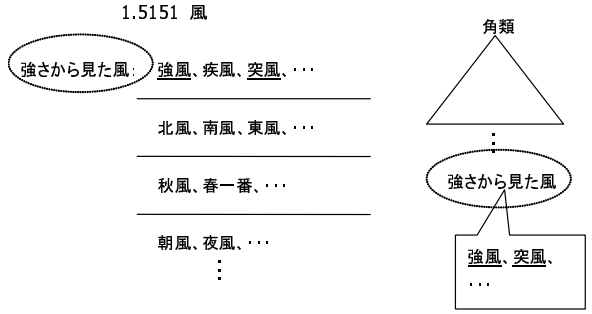

図6：国類の分類観点推測

4　仮想シソーラスの語の検索

仮想シソーラスから単語を検索するには、他シソーラスから目的の語を検索し、その語を含む他シソーラスの末端ノードが岩類のどの末端ノードに接ぎ木されるかを調べれば良い。構築時とは逆に、他シソーラスの末端ノードの各語を岩類から検索し、一致語数の割合が一定以上の岩類ノードがあれば、その岩類の末端ノードの下に他シソーラスの末端ノードが接ぎ木されることがわかる。

図7に検索の一例を示す。“犬”という単語を検索すると、他シソーラスの末端ノード“獣類のいろいろ”“密偵”などの下に存在することがわかる。ここで“獣類のいろいろ”“密偵”の各語を岩類で検索すると、それぞれ岩類の“獣”“スパイ”というノードの下に接ぎ木されることがわかる。

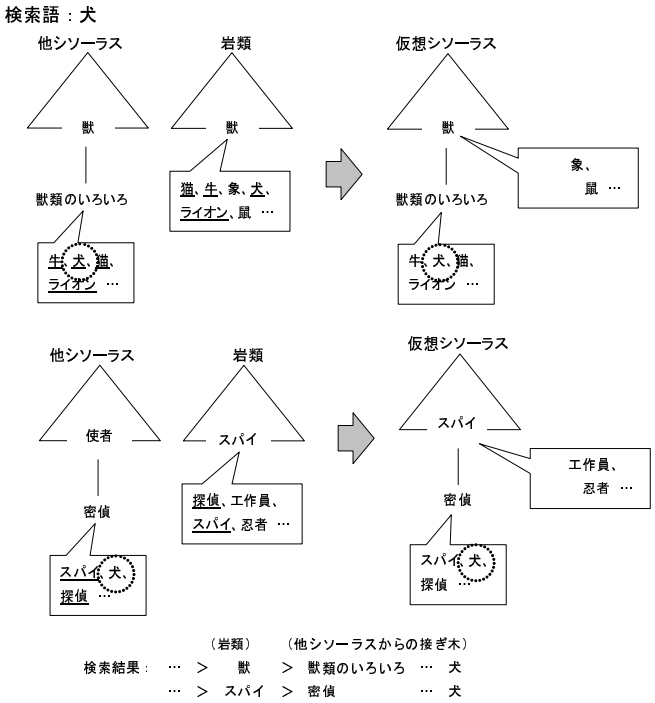

図7：仮想シソーラスの語の検索

5　自動化ツールの作成

以上の手法を用いて、実際に仮想シソーラスを構築し表示・検索するツール (cgi) を作成した。図8、図9にその実行例を示す。

図8の例について説明する。左上のフレームには岩類のツリー構造が表示され、細分化したいノード（この例では“風”）を選択すると、左下のフレームにそのノード内の語（細分化前）が、右上のフレームには他シソーラスを使って細分化された後のノードが表示される。さらに右上のフレーム内のノードを選択すると、右下のフレームにそのノード内の語が表示される。

また、不要なノードを判別する一致語数の割合の基準値については、右上のフレームから自由に変更することができる。

図 8：cgi 実行例 (表示)

図 9：cgi 実行例 (検索)

図 9 は、検索ツールを用いて文字列“犬”を仮想シソーラスから検索した例である。詳しく見たいノードを選択することで、そのノードの詳細を表示 (図 8 を参照) することができる。

6　おわりに

既存の複数のシソーラスを組み合わせ、それぞれの長所を合わせ持った仮想シソーラスを構築する方法を提案した。また、多次元シソーラス作成補助ツールとして、末端ノードの対応付けに使う一致語数の割合の基準値を変えながら仮想シソーラスを表示・検索する cgi を作成した。

今回は 3 つのシソーラスを用いたが、今後有用なシソーラスが登場した場合、追加で組み込むことで更にシソーラスの語彙・分類観点の増強を図ることができる。

参考文献

[1] 川村、片桐、宮崎：語を種々の観点から分類した多次元シソーラス、電子情報通信学会技術報告、NLC94-48(1995)

[2] 森田、宮崎：連想型多次元シソーラスとその意味解析への適用性、言語処理学会第 12 回年次大会、A4-2(2006)

[3] 池原、宮崎、ほか 6 名：日本語語彙大系、岩波書店 (1997)

[4] 大野、浜西：角川類語新辞典、角川書店 (1981)

[5] 国立国語研究所：分類語彙表、秀英出版 (1964)