情 報 学 基 礎 42－4
（1996. 7. 25）

複数の言語資源からのシソーラスの構築

松本 裕治，須藤 茂†，中山 拓也，平尾 努
{matsu, sigeru-s, takuya-n, tutomu-h}@is.aist-nara.ac.jp
奈良先端科学技術大学院大学 情報科学研究科
〒630-01 奈良県生駒市高山町 8916-5
†現在，日本電信電話株式会社

語の意味的な類似性を表現する辞書として，シソーラスは自然言語処理における重要な要素の一つである．本稿では，一般の国語辞典およびコーパスから得られる語の係受け関係を利用したシソーラス構築について述べる．コーパスからの係受けや共起関係に基づいて語の意味的類似度を測り，クラスタリング手法を用いてシソーラス構築を行なう研究が多く行なわれているが，本稿では，語が表す概念間の上位下位関係を抽出することを目的とする．機械可読辞書の定義文から抽出された語の上位下位の階層関係とコーパスからの共起関係によって統計的に得られる語の意味的類似性および語の使用範囲の多様性から予測される概念の包含関係を組み合わせることにより語の階層構造を漸進的に構築する方法を提案する．

THESAURUS CONSTRUCTION FROM MUTIPLE LANGUAGE RESOURCES

Yuji Matsumoto, Sigeru Sudo, Takuya Nakayama, Tsutomu Hirao

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara, 630-01 Japan

A thesaurus is a dictionary to represent sematical similarity and relationship between words. It plays an important role in language processing. This paper discusses a method to construct a thesaurus from the information obtained from a Japanese dictionary and depedency ralation of words taken from a language corpus. As opposed to the word classe extraction using clustering methods on word collocation, our method tries to obtain an IS-A hierarchy composed by word concepts. Hierarchical relationship between words is extracted from the definition sentences in a dictionary and is augmented by the information obtained from corpus analysis, such as the similarity and expected inclusion relationship between words.

1 はじめに

言語処理のための基本的な意味情報源としてシソーラスの担う役割は大きい．国内では，国立国語研究所の分類語彙表[中野 96]が電子データとして入手可能であり，広く使われている．公開版の分類語彙表の電子データの語数は3万数千であるが，最新の増補版では8万7千を数える．この辞書では，語の意味に従って，体，用，相それぞれ6階層の木構造の葉の部分に日本語の単語が配置されている．

何万，何十万という語を含むシソーラスの構築には多大の労力が必要である．また，語の意味や用法には通時的な変化や分野依存性が強いことが多いため，各種言語資源からのシソーラス自動構築の研究が行なわれている．次章では，代表的手法である機械可読辞書からのシソーラス構築，および，コーパスにおける語の共起関係を利用したシソーラス構築について概観する．

本稿では，語の意味関係の中でも概念間の上位下位関係(IS-A関係)に注目し，語による階層構造を抽出することを目的とする．このように語の上位下位関係を表現したシソーラスの代表例としてWordNet[Miller 90]がある．本稿では，日本語において同様のシソーラスを構築する際の方法論として，いかにして既存の言語資源を利用するかという問題に焦点を当てる．

2 言語資源からのシソーラス構築

本節では，電子化された言語資源からのシソーラスの自動構築に関する研究を概観する．この種の研究には辞書およびコーパスが利用されることが多いので，それぞれに分けて現状と問題点を見る．

2.1 辞書の定義文からのシソーラス構築

機械可読辞書の定義文を解析することによって，その語の意味的な上位語を取り出すことができる．特に，辞書の定義文には定型的な文が使われることが多く，深い言語解析をせずとも，表層的なパターン照合によって上位語を取り出すことができる場合が多い．

先駆的な研究としては，Amsler[Amsler 81] やChodorow[Chodorow 85] らが辞書の定義文からの上位語抽出のためのパターン照合手法を提案している．日本語の国語辞典についても，同様の手法による鶴丸らの研究がある[鶴丸 91]．このような手法による上位語抽出にはいくつかの問題点がある．すなわち，

- 定義文において上位語として使われる語は，定義される語と概念的に近い直接の上位語ではなく，一般性の高い語が使われる傾向がある．したがって，取り出される階層構造は平坦なものになりがちである．
- 上位語が意味的に多義の場合に，どれが正しい意味であるかを判別できない．
- 階層構造の上位に属する語の定義がループをなすことがある．
- 定義文の中心語が必ずしも意味的な上位語になるとは限らない(～の一種，～の総称，～であること，等)．したがって，上位語を取り出すための何らかの方法あるいはヒューリスティックスが必要である．

複数の機械可読辞書を用いることによって精度が向上したという報告もある[Ide 93]が，上の多くの問題が解決されるわけではない．

多義性の問題を除けば，辞書からは間違った階層関係が取り出される可能性は低い．しかし，それらが概念的に直接，あるいは，近い上位下位関係を表しているかという適切性，および，概念間に成り立っている階層関係をどの程度取り出し得るかとい再現性を考えると，両者とも高いとは言えない．鶴丸の研究でも，辞書から取り出された階層構造は，比較的平坦でありかつ互いに関係付けられず独立した多くの階層構造の集合になったことが報告されている．

2.2 コーパスにおける共起関係からのシソーラス構築

コーパスにおける語と語の共起関係に基づいて語の間の意味的な距離を求め，それに基づいてシソーラスを構築する研究が数多く行なわれている．共起関係としては，語間の近接性あるいは格関係などの文法的係受けのいずれかが用いられるが，動詞の格関係(例えば動詞とその目的語)に基づいて名詞の集合のクラスタリングを行なってた研究が多い．

Hindle[Hindle 90]は，名詞のクラス分けを行なうために相互情報量を用いて語の類似性を定義している．その他，Pereira[Pereira 93], Tokunaga[Tokunaga 95], Li[Li & Abe 96], 平岡[平岡 94]らも動詞とその格要素の関係に基づいて名詞をクラスタリングする方法を提案している．クラスタリングの方法は，名詞の集合を意味的に近い語が同じクラスに属するように分割していく方向と意味的に近い語を同じクラスにまとめ上げていく方向があるが，いずれにしても得られる構造は，分類語彙表のように構造の最下位に語が配置された木構造である．

コーパスからの情報に基づいた語のクラス分けは，

コーパスの性質を反映しているという点で分野依存性に強く，同分野のテキストを解析するために有用な資源となる可能性がある．しかし，一方，次のような問題点も存在する．

- コーパスに出現頻度の低い語に対しては正確な情報が得られない．
- 出現頻度が高くても，特殊な言い回しあるいは固定的な言い回しでしか現れない語は概念的には不自然な語と同クラスに入ることがある．
- でき上がった構造は概念間の上位下位などの階層構造を必ずしも表していない．

コーパスを利用した語のクラス分けはコーパスの性質を反映しているため，同じ分野のテキストの係受けの曖昧性解消や格フレーム記述には適しているかも知れない．しかし，得られたシソーラスの階層構造は対象とした特定の共起関係に引きずられることが多く，また，客観的な意味付けを与えることが困難である．

2.3 シソーラスの自動獲得 vs 半自動獲得

上に挙げた研究の多くは，電子化された言語資源から自動的にシソーラスを構築する研究である．しかし，どの研究も小規模あるいは特定の目的にのみ使えるデータしか得られない．現在利用可能な言語資源を用いて，ある意味で汎用のシソーラスを構築することは極めて困難である．特に，WordNet のように概念の正確な階層構造を得ることは，コーパスからの情報だけでは不可能である．

概念の階層関係を取り出すために辞書の記述は有用な情報であるが，上に述べたようにそれだけでは不足である．本稿では辞書からの情報とコーパスからの共起関係から得られる情報を組み合わせて利用することによりシソーラス構築を試みる．もう一つの重要な視点は，シソーラスの構築に人間の介入を許す半自動構築を考えた時にこれらの資源がどのように利用され得るかということである．完全な自動獲得が難しい場合，人間の介入および個人差をどのように最小限に限定し，かつ，効率よく行なうことができるかを考慮する必要がある．

本稿では，部分的に記述された階層構造 (辞書の解析により機械的に取り出されたものでも人手によって記述されたものでもよい) およびコーパスから得られる動詞とその「を」格に現れる名詞との共起関係から統計的に得られた類似度および包含可能性関係に基づいて，シソーラスを漸進的に獲得する方法について論ずる．

3 辞書とコーパスからの漸進的シソーラス構築

辞書の定義文から取り出される概念間の上位下位関係とコーパスにおける共起関係から得られる情報を組み合わせることを考える．ただし，両者からも完全な階層関係を得るための十分な情報を望むことはできず，また，組み合わせ方も明確な方法がある訳ではない．まず，辞書とコーパスからの情報の抽出について述べる．

3.1 辞書の定義文からの上位語の抽出

ここで用いる方法は，機械可読辞書に対して従来行なわれたきた方法と同じものを採用する．各名詞の定義文の第一文を形態素解析し，その中心語 (日本語の場合は最後に現れる名詞) を上位語とする．ただし，辞書に現れる特殊な表現 (～の一種，～の総称，～の略，等) についてはヒューリスティック規則を作成し，中心語の候補となるものをその直前の表現から求めている [須藤 96]．特に，定義文が単一の語からなる場合は，それを見出し語と同義語と判断する．辞書としては，大辞林 [松村 88] を題材とした．例えば，次のような記述からは太字で示した語がそれぞれ上位語および同義語として取り出される．

銀行: 預金の受入、… などを主たる業務とする**金融機関**。

財団: 「**財団法人**」の略。

機械可読辞書からはこのような情報を自動的に得ることができる．しかし，本稿では，このような情報は単に一つの資源 (人間を含む) から得られる確度の高い階層関係であるとみなし，辞書から得られる関係もある程度取捨選択されるものとする．

3.2 コーパス中の動詞と名詞の共起関係に基づく名詞の類似度計算

コーパスとして毎日新聞の記事データ (93,94 年版) を用いた．全記事の本文に形態素解析を施し，格助詞「を」を伴う名詞とその後に現れる動詞 (サ変名詞を含む) を機械的に抽出した．抽出項目の総数は約 130 万である．ただし，実験では出現頻度が 20 回以上の名詞 3162 個のみを対象とした．記事の分野を限定することも考えられるが，データ量を重視して，本実験では，2 年間の全記事を用いた．

各名詞に対して共起する動詞の頻度ベクトルを考える．これを確率分布と見ると，名詞間の類似度を定義するために確率分布間の近さを表す尺度を利用するこ

とができる．ただし，すべての動詞を表層形のままで利用すると，約2000次元のベクトルとなり，頻度の低い動詞が多数存在するので，動詞をクラス分けすることを考えた．動詞のクラス分けも分野などを考慮することが望ましいが，実験では分類語彙表の用の部の下位から第2階層までの語が意味的に類似と判定し，同じ下部の階層に属する動詞を同一視した．得られた動詞クラスの数は151である．

名詞 n に対する動詞クラス $v(v \in V, V$ は動詞クラスの集合, $|V| = 151)$ の共起頻度を $f(n,v)$ とする．共起頻度の低い(0回を含む)動詞クラスへの補正および全体の頻度の違いを吸収するために次のように名詞 n に対する確率分布の推定値 $p_n^*(v)$ を定義する．ここに S は出現頻度補正のための標準値 $(S = 5000$ とした$)$，N は n の総出現頻度である．また，$r = S/N$とする．

$$p_n^*(v) = \frac{r \cdot f(n,v) + 1}{S + |V|}$$

確率分布の距離を測る尺度として，次の Kullback-Leibler 情報量が使われることが多い．すなわち，2つの名詞 n, m に対する確率分布を $p_n^*(v), p_m^*(v)$ とするとき，次の式で与えられる非負数によって名詞間の類似度を定義する．

$$I(p_n^* \parallel p_m^*) = \sum_{v \in V} p_n^*(v) \log \frac{p_n^*(v)}{p_m^*(v)}$$

Kullback-Leibler 情報量は，2つの確率分布に対して対称ではないので，本稿では，双方の確率分布から見た Kullback-Leibler 情報量の和である Jeffery 情報量 [稲垣 90] を用いることとした．

$$J(p_n^* \parallel p_m^*) = \sum_{v \in V} (p_n^*(v) - p_m^*(v)) \log \frac{p_n^*(v)}{p_m^*(v)}$$

Jeffery 情報量も非負数であり，その値が小さいほど2つの名詞の類似度が高いと解釈する．また，同一の確率分布による値は0となる．類似度の高い名詞はシソーラス中で近接して存在することが予測されるが，概念の上位下位の記述を目標とする場合には，2つの近接する名詞が兄弟関係か親子関係のいずれの関係を持つのか，また，親子関係としてもどちらが上位の概念と考えるのがふさわしいのかを判断する必要がある．

辞書あるいはその他の情報からそれらが上位下位の関係にあることがわかっている場合は問題がないが，そうでない場合は，コーパス中における出現の仕方によって，関係を予測する必要がある．ここでは，単純な考え方であるが，下位の語は上位の語より特殊な語であるため用法が限定されていると仮定し，限定の度合をそれぞれの方向から見た Kullback-Leibler 情報量の差によって予測することにした．$I(p_n^* \parallel p_m^*)$ と $I(p_m^* \parallel p_n^*)$ の値は，n が m に比べてより限定的な (ピークをもった) 分布である時に後者の方が大きい値を与える．したがって，これらの量に有意差がある場合 (後者の方が有意に大の場合．後で示す実験では一方が他方の1.5倍以上の時に有意差があるとした)，n の方が m より限定された概念であるとみなし，n を下位概念とする解釈を優先する．有意差が認められない場合は，兄弟関係とみなす解釈を優先する．

3.3 漸進的シソーラス構築

本稿では，全自動のシソーラス構築を目指すのではなく，人間の介入を認める．現実的な利用としては，シソーラス構築を支援するシステムとして利用することを考えている．以下に述べる手順は，自動のシソーラス構築法として見ることもできるが，本稿の目的は人間にとっても自然な意味を反映したシソーラスの構築であり，全自動による構築を期待するのは現実的ではない．提案する手法は以下の処理の繰り返しである．辞書の定義文からの階層関係とコーパスから予測される階層関係をすべて同時に考慮して何らかの最適解を求めることも考えられるが，与えられたすべてのデータに対して無矛盾な解が得られる保証はなく，計算量の上からも現実的ではない．

1. シソーラスの作成を開始するための単語の集合 (通常は小さい集合) を選択し，辞書から階層関係を得る．最終的な語の集合を Wとする．
2. Wから適当な語 wを選び，Jeffery 情報量の小さい語を事前に決められた類似度以内でかつ決められた数だけ求める．選択された語の集合を Cとする．
3. Cのそれぞれの語の辞書の定義文から得られる上位語をシソーラスに追加する．
4. Cのそれぞれの語と wおよびその上位下位の語との両方向の Kullback-Leibler 情報量を求め，既に構築されたシソーラス中の語との関係を予測する．矛盾の生じない限り，シソーラスに追加する．
5. 更新されたシソーラスをもとに 2. からの手続きを繰り返す．

この手続き中の 4. で矛盾が生じた場合には，1) その語のシソーラスへの追加を中止する，2) 処理を中断して人間の介入を求める，などの処理が考えられる．まだ実現はしていないが，本研究では，後者の利用形態を想定し，人間による修正を伴いながら，辞書とコー

パスからの情報を利用して漸進的にシソーラス構築を行なう環境を想定している．

4　シソーラス構築の実験例

上記手法の有効性を確認するために，小規模の実験を行なった．その具体例を示す．実験では，前に述べたように辞書の定義文としては大辞林を用い，コーパスとして毎日新聞2年分(93,94年)から「を」格によって結ばれた(と考えられる)名詞と動詞の対を取り出してJeffery情報量を計算した．なお，サ変名詞は動作概念を表すので，普通名詞と階層関係をもつことは少ないと考え，普通名詞のみを対称とした．また，手続きの2.で選択される語は類似度が2.5以下かつ10個以内に限定した．なお，辞書記述における多義性の影響を限定するため，分類語彙表を利用して上位から2階層目で異なる部分木に入る上位語は考慮の対称から外した．

実験例

「会社」という語1つからなる集合を初期集合とし，前章の手順を適用した.「会社」の辞書記述からは「会社 < もの」[1]という関係しか得られなかった．選択された語と類似度は次の通りである．

[会社] 出現数 = 468

研究所	1.55791671184922
銀行	1.70691573603657
子会社	1.71507679762516
センター	1.88563520487119
会	1.99530186892006
財団	2.02208263738301
機構	2.05958875338957
協会	2.06466792823837
機関	2.19747126434552
所	2.34697836265047

これらの語をもとに辞書の定義文から会社 < もの，子会社 < 会社，銀行 < 金融機関，金融機関 < 機関，センター < 機関，機構< 仕組み，などの関係が得られる．

次に，コーパスからの類似度と特殊性の判定より，研究所 < 会社，財団 < 基金，財団 < 協会，機関 < 機構，協会< 機関，などが推定される．

手続きが一巡した後の階層関係を「機構」中心に図示したものが図1,「会社」を中心に示したものが図2で

[1] $A < B$はBがAの上位語であることを意味する．

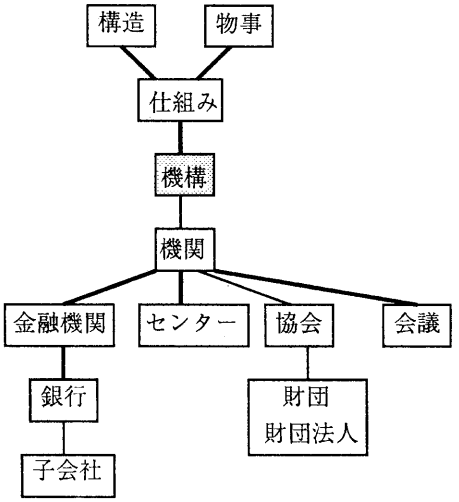

図 1:「機構」を中心とする階層関係

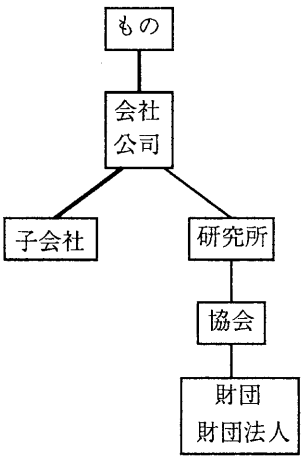

図 2:「会社」を中心とする階層関係

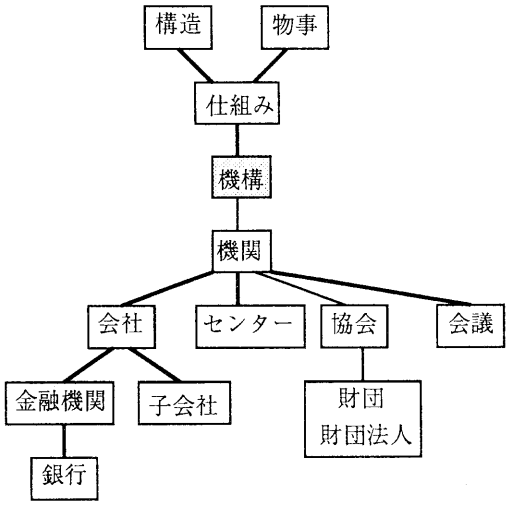

図 3: 修正された階層関係

ある．太線は辞書の定義文から得られた階層関係，細線はコーパスから予測された階層関係である．

コーパスからの情報によると「銀行」に最も類似の語は「会社」であり，Kullback-Leibler 情報量も有意差がない．「銀行」の上位語および関連語と「会社」との間の Kullback-Leibler 情報量を考慮すると，「機関」の上位あるいは下位に矛盾なく「会社」を配置することができなくなる．ここで人間が介入するすることにより，「会社」を「機関」と「金融機関」の間に配置し，不正確と思われる，「子会社」と「銀行」の間のリンク，および，「研究所」と「会社」のリンクを切ることにより図 3 のような修正された階層構造を得ることができる．

5 おわりに

自然言語処理のための資源としてだけではなく，知識ベースとしても概念の階層関係を記述したシソーラスは有用な資源である．様々な言語資源を解析することによってシソーラス構築を目指した研究があるが，本稿では，機械可読辞書およびコーパスから得られる情報を利用したシソーラス構築について述べた．本稿で述べた手法は，システムとしてはまだ部分的にしか実現されていないので，ここでは小規模な実験例を提示するに留まった．

本研究は，複数の言語資源を統合的に利用する方法を探ることを最初の動機として始まった．現在入手可能な資源や言語解析システムからシソーラスを自動的に構築することは極めて困難であり，いかに人間の介入を限定するか，あるいは，人間の介入を効率よく行なうことができるかを考察する方が建設的である．本稿ではその一つの試みを提案した．

本稿で述べた方法は，計画段階の内容であり，今後に残された問題は多い．そもそも，階層構造の概念ラベルをすべて単語だけで表現することも見直すべきである．WordNet のように語の集合として概念を表す方法やそれを本枠組に取り込む方法を考える必要がある．また，語の多義性の問題や人間の能力を最大限引き出し，かつ，介入を効率よく行なうためのインターフェースについても研究する必要がある．

謝辞: 本研究では，CD-毎日新聞 93,94 年版を利用した．新聞記事データの研究利用許諾をいただいた毎日新聞社に感謝する．

参考文献

[Amsler 81] Amsler, R., "A Taxonomy for English Nouns and Verbs," Proc. 19th ACL, pp.133-138, 1981.

[Chodorow 85] Chodorow, M., et al., "Extracting Semantic Hierarchies from a Large On-Line Dictionary," Proc. 23th ACL, pp.299-304, 1985.

[Hindle 90] Hindle, D., "Noun Classification from Predicate-argument Structure," Proc. 28th ACL, pp.268-275, 1990.

[平岡 94] 平岡冠二，松本裕治，「コーパスからの動詞の格フレーム獲得と名詞のクラスタリング」，情報処理学会自然言語処理研究報告, 94-NL-104, pp.79-86, 1994.

[Ide 93] Ide, N. and Véronis, J., "Extracting Knowledge Bases from Machine-Readable Dictionaries: Have We Wasted Our Time?" Proc. KB&KS, pp.257-266, 1993.

[稲垣 90] 稲垣宣生：数理統計学，裳華房，1990.

[Li & Abe 96] Li, H. and Abe, N., "Clustering Words with the MDL Principle," Proc. 16th COLING, to appear, 1996.

[松村 88] 松村明 編，「大辞林」，三省堂，1988.

[Miller 90] Miller, G., et al., "Fibe Papers on WordNet," CSL Report 43, Princeton University, 1990.

[中野 96] 中野洋，「分類語彙表」形式による語彙分類表 (増補版)，国立国語研究所，1996.

[Pereira 93] Pereira, F., et al., "Distributional Clustering of English Words, Proc. 31st ACL, pp.183-190, 1993.

[須藤 96] 須藤茂，「複数の言語資源を用いたシソーラスの構築」，奈良先端大情報科学研究科修士論文，NAIST-IS-MT9451060, 1996.

[Tokunaga 95] Tokunaga, T., et al., "Automatic Thesaurus Construction Based on Grammatical Relation," Proc. 14th IJCAI, Vol.2, pp.1308-1313, 1995.

[鶴丸 91] 鶴丸弘昭，他，「国語辞典を用いたシソーラスの作成について」．情報処理学会自然言語処理研究報告, 91-NL-83, pp.121-128, 1991.