Vol. 37 No. 12 情報処理学会論文誌 Dec. 1996

# コーパスに基づくシソーラス——統計情報を用いた既存のシソーラスへの未知語の配置

浦本直彦†

本論文では，コーパスに基づくシソーラスを構築するための基礎として，既存の中規模のシソーラスとコーパスを用いて，シソーラスを拡張する手法について述べる．シソーラス上にない単語に対して，その単語がシソーラスのどの部分に配置される可能性が高いかをコーパスから抽出した統計情報を用いて決定する．シソーラスの分類基準（視点）を自動的に獲得することで，効率良く単語の位置を推定することが可能である．これらの知識を用いて，拡張されたシソーラス上での位置，上位語，単語間の類似度などを計算する関数群を提供するためのシステムを作成した．

# Corpus-based Thesaurus — Positioning Words in Exsiting Thesaurus Using Statistical Information from a Corpus

NAOHIKO URAMOTO†

This paper describes development of a corpus-based thesaurus system. For the purpose, a method for positioning unknown words in an existing thesaurus is proposed. A likely area of the thesaurus for an unknown word is estimated by integrating the human intuition buried in the thesaurus and statistical data extracted from the corpus. To overcome the problem of data sparseness, distinguishing features called "viewpoints" of each node are extracted automatically and used to calculate the similarity between the unknown word and a word in the thesaurus. The results of an experiment confirm the contribution of the viewpoints to the positioning task. By using some functions for accessing the thesaurus with viewpoints, users can get information for words in the thesaurus including unknown words.

## 1. はじめに

単語間の上位下位関係や同義関係を記述したシソーラスは，自然言語処理分野における最も重要な知識源のひとつである．たとえば，情報検索においては，ユーザが入力した検索式を拡張するのに使われる．また，単語間の意味的な距離を計算するために，多義性の解消や機械翻訳といった多くのシステムにおいて利用されている．英語に関しては，Rogetのシソーラスや WordNet[3]，日本語においては，分類語彙表などを用いて，これまで数多くの研究がなされている．

しかし，シソーラスはもともと，工学的利用を念頭に入れずに作られたものが多く，シソーラスを自然言語処理システムにおける利用という面から見た場合，次の3つの問題がある．

第1に，既存のシソーラスでは，語彙が不十分である．また一般的な語を中心に構成されており，特定分野のシソーラスは存在しないことが多い．

第2に，シソーラスを構成する際に，シソーラス作成者は，内省と言語資料を基に，単語をいくつかの基準に従って配置したはずである．それらの基準は，シソーラスを工学的に利用する際に，非常に有用であると考えられるが，それらの情報は，シソーラス上には明示的に記述されていない．逆に，大量のテキストから抽出した情報に基づいて統計的な手法を用いてシソーラスを自動生成する研究もなされているが，構築されたシソーラスが人間の直観に合わないことが多いという問題が生じている．

第3に，既存のシソーラスは，たとえば，単語間の距離（類似度）を計算するのに用いられるが，その階層の深さのバランスがとれていなかったり，その体系に恣意的な部分が存在するため人間の直観に合わないことがある．たとえば，生物的学的な分類は，階層の数が長くなることが多いし，抽象的な単語は比較的浅い階層構造上におかれることが多い．また，シソーラ

† 日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research, Tokyo Research Laboratory

スを適用する単語が含まれるテキストの分野によって，階層の体系が微妙に変わることがある．

これらの問題を解決するために，著者らは，コアとなる中規模のシソーラスと，分野依存のテキストから抽出した統計情報を組み合わせた，コーパスに基づいたシソーラスシステム（Corpus-based Thesaurus System）の構築を行っている．このシステムでは，ユーザから見て，すべての単語がシソーラス上に規則正しく配置されている必要はなく，コアとなるシソーラスと統計情報を用いて，動的に単語間の関係や類似度が計算される．ユーザは，あらかじめ定義された関数（API）を通して，シソーラスシステムに対し入出力を行う．コアシソーラスには，汎用のものを用い，分野依存のテキストコーパスからの統計情報を用いて，コアシソーラスを補強したり，アクセス関数を通じて，単語間の関係を得ることが可能である．

本論文では，コーパスに基づくシソーラスを構築するための基礎として，既存の中規模のシソーラスとコーパスを用いてシソーラスを拡張する手法について述べる．具体的には，シソーラス上にない単語に対して，その単語がシソーラスのどの部分に配置されるのかを決定するというタスクである．本論文では，これらの単語は，テキストコーパス中には複数回出現するが，あらかじめ語義が分かっていないという点で，未知語と呼ぶ．

未知語のシソーラスへの配置問題は，単語の語義の曖昧性解消問題[1),2),4)~6)]の変形と見なすことができるが，その違いは，解空間の広さである．従来の多義性解消では，あらかじめ語義の候補は分かっているので（たとえば，bankの語義は「土手」か「銀行」である），その語義のうちのどれか（一般的には，たかだか十数個）を決定することになる．ところが，未知語の場合，あらかじめ語義は分かっていないので，解の候補はシソーラス上のすべての単語であり，シソーラス上での正確な位置を示すのは非常に困難である．そこで，本論文では，未知語がシソーラスのどの部分に属す可能性が一番高いかを推定する．

また，シソーラス作成者はある分類基準に従って，単語をシソーラス上に配置したはずであり．未知語のシソーラス上での位置を推定するためには，シソーラスを構築するために用いられた基準を用いるということが考えられる．この基準を視点と呼ぶ．この視点は，たとえば，「～が飛ぶ」「～が着陸する」といえるものは，「航空機」に分類するといった，助詞などをともなう単語間の係り受け関係として抽出される．本論文では，シソーラスの各ノードに対して，統計情報を用いて視点を明示的に記述し，未知語の配置のための手がかりとする．

| | |
|---|---|
| 具体物 (Physical Object) | 現象 (Phenomenon) |
| 有意志体 (Creature) | 関係 (Relation) |
| 抽象物 (Abstract Object) | 時 (Time) |
| 方法 (Method) | 場所 (Location) |
| 活動 (Action) | 空間 (Space) |
| 属性 (Attribute) | 単位 (Unit) |
| 状態 (State) | 操作 (Operation) |
| 力 (Force) | |

**図1** ISAMAPの最上位の分類項目
Fig. 1 Top categories of ISAMAP.

## 2. シソーラスとコーパスから抽出した統計データ

本章では，本論文で用いるシソーラスと，統計データについて述べる．シソーラスとして，東工大の田中らが作成したISAMAP[7)]を用いる．ISAMAPは，約4,000の一般的な名詞を，上位/下位関係の観点から階層的に配置したものである．いくつかの単語は複数の位置に分類されている．**図1**に，ISAMAPの最上位の分類項目を示す．

シソーラス上での分類のための視点を抽出し，未知語をシソーラス上に配置するために，1993年の日本経済新聞記事から抽出した助詞などの単語間の関係を表す語をともなった単語の2項関係を用いる．記事データは，形態素解析され，次のような形式で記述されている．

$oc$(単語1, *rel*, 単語2) $= n$

これは，単語1と単語2が，relという関係を表す語（関係マーカと呼ぶ）を介して係り受け関係にあり，この係り受け関係が記事データに$n$回出現していることを示している．たとえば，$oc$(犬, が, ほえる)は，「犬がほえる」という係り受け関係の出現を示す．関係マーカとしては，「が」「を」などの格助詞と，「な」「い」のような，連体修飾関係を指示する語を用いる☆．**表1**に，記事データから抽出した単語間の関係のサイズを示す．

従来の研究として，Yarowskyは，あらかじめ語義が分かっている単語の多義性を解消するのに，それらの単語とある範囲内で共起する単語の集合を用いている[5)]．本論文では，単語の集合ではなく，単語間の係

☆ 本論文では，関係マーカとしては，表層的なレベルで抽出できるような格情報を中心に設定した．

表1 統計データのサイズ
Table 1 Size of statistical data.

| 関係マーカ | 異なり語数 | 延べ頻度 |
|---|---|---|
| が | 394.887 | 817,030 |
| を | 483,400 | 1,210,581 |
| へ | 18,564 | 53,876 |
| に | 451,986 | 1,114,877 |
| で | 225,247 | 61,4619 |
| と | 176,738 | 570,475 |
| な | 78,079 | 569,837 |
| い | 51,001 | 881,255 |
| 合計 | 1,879,902 | 5,832,550 |

り受け関係を用いているが，それには，次の2つの理由がある．

まず，あらかじめいくつかの語義が分かっている単語の場合，語義の候補はもちろん語義の数だけあり，この中から正しい語義を決定するわけであるが，未知語の場合，その語の意味の候補はシソーラス上の単語すべて（ISAMAPでは約4,000）である．語義の推定を妨げるノイズを除去するために，より制限の強い関係を用いている．

第2の理由は，格関係のような関係マーカによって明示される単語間の関係は，単語分類の視点として用いることができるためである．たとえば，シソーラス上で，「飛行機」と「船」が並列ノードのとき，この2つの単語およびそれらの下位語を分類するのは，「飛行機が飛ぶ」というが「船は飛ぶ」とはいわない，という現象である．つまり「～が飛ぶ」という（係り受け）関係は，「飛行機」を分類するという視点のひとつになり得る．

## 3. 未知語のISAMAP上の位置の推定

本章では，未知語が，ISAMAP上のどの部分に位置するかを推定する手法の流れについて述べる．処理は，次の3つに分けられる．

**Step 1:** ISAMAP上の各単語（ノード）に対して，視点を抽出する．

**Step 2:** 入力となる未知語に対して，その単語が属する可能性のあるISAMAP上での部分（エリア）を求める．

**Step 3:** 各候補エリアを評価し，最も優先度の高いものを未知語の属す部分とする．

### 3.1 基本的な考え方

処理の基本的な考え方は単純である．まず，入力となる未知語に対して．コーパスから，その語を含む関係 $oc$（前章参照）を検索する．これらの単語間の関係と，ISAMAP上の各単語ノードに関する関係を用いて，両者の類似度を計算する．ある一定の閾値を超える類似度を持つノードに，マークを付ける．図2(a)に，未知語に対し類似性が高いとマークされた単語の分布の例を示す．この図の中で，黒丸がマークされた単語である．単純に単語間の類似度を計算する手法では，シソーラス上に分散した単語がマークされてしまい，入力単語がシソーラスのどの位置に属すべきかを決定することができない．これには，たまたま特定の用法が似ていたとか，なんらかのノイズが混入したといった理由が考えられる．そこで，これらのノイズの影響をできるだけ避けるために，次のような規則を適用する．

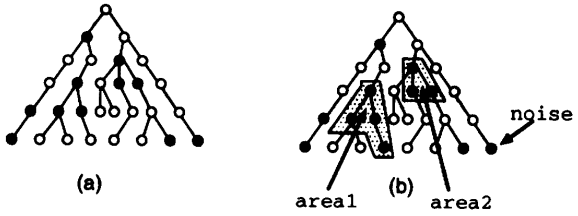


図2 シソーラス上でマークされた単語ノード
Fig. 2 Marked nodes in the thesaurus.

(1) シソーラス上の単語を用いるのではなく，マークされた単語の連結された集合（エリアと呼ぶ）を用いる．図2(b)のarea1，area2がそれにあたる．より大きなエリアほど入力語がそこに属する可能性が高い．

(2) より特定的な単語，つまり，シソーラスでより下位に属する語を優先する．図2(b)の場合，area1がarea2より優先される．

(3) 入力単語が，シソーラス上の単語の視点を含む場合，その単語を含むエリアは優先される．たとえば，「飛行機」という単語とその下位語の視点として，「～が飛ぶ」という係り受け関係が抽出された場合，未知語 $W$ に「$W$ が飛ぶ」という係り受け関係があるのなら，$W$ は，「飛行機」と密接な関係があるといえる．

### 3.2 視点の抽出

視点とは，シソーラス上のそれぞれの単語が，他の単語に対してどのような差異を持っているか，また，その単語がどのような基準でシソーラスの特定の位置を占めているのかに対して，何らかの尺度を与える情報である．この情報は，本論文では，単語間の関係としてとらえられる．シソーラスの各ノード $w$（ISAMAPの場合，ノードは単語に等しい）に対して，視点は，次のように関係マーカ，視点となる単語，および視点となる単語が $w$ にどのように係るのか（視点単語が $w$ に係るときは“modifier”，係られる時は“head”）

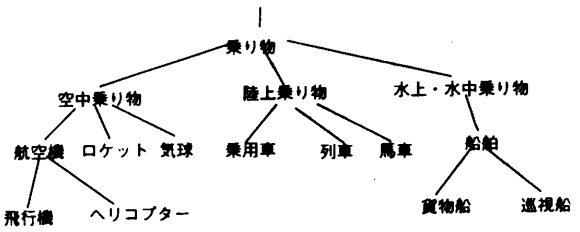


図3 シソーラスの例
Fig. 3 Example of ISAMAP.

表2 抽出された関係の例
Table 2 Example of extracted relationships.

| 単語 | マーカ | 頻度 |
|---|---|---|
| 使う | を | 10 |
| 鳥 | で | 5 |
| 救助する | に | 3 |
| 飛ぶ | が | 2 |
| つり上げる | で | 2 |
| チャーターする | を | 2 |
| 含む | を | 2 |
| 甲板 | で | 2 |
| 出動 | を | 2 |
| 上空 | で | 2 |

の組で記述される．また，単語 $w$ に対する視点は複数あり，$w$ に対する視点集合として定義される

視点集合 $(w)$
$=\{(rel_1, word_1, pos_1), (rel_2, word_2, pos_2), \ldots\}$

このような属性は，シソーラスを構築する際，作成者が内省的に，あるいは明示された基準として用いていると考えられる．

視点抽出の流れは次のようになる．**図3** にISAMAPの一部を示すが，この部分に含まれる単語の視点を抽出することを考える．たとえば，「ヘリコプター」という単語に対して，テキストコーパスから獲得した単語間の係り受け関係を用いる．ちなみに，この単語は，コーパス中に 131 回出現している．**表2** に，抽出された係り受け関係の例を示す．

各々の関係に対して，同じ関係を持つシソーラスに含まれる単語を調べる．たとえば，最も頻度の高い(を，使う，"head") という関係の場合，「ヘリコプター」を含む，シソーラス中の 385 個の単語が同じ係り受け関係を持つ．これに対し，(が，飛ぶ) というパターンでは，「ヘリコプター」「飛行機」だけである．図3 に示すようにこれらの単語は，兄弟関係がある．ある単語の視点は，多くの場合，下位にも継承される性質を持つので，(が，飛ぶ，"head") という関係は，「飛行機」「ヘリコプター」の視点だと考えられる．

その関係が，どのくらい単語を差別化するかという尺度 typicalness を次のように定義する．単語 $nd$ に対する関係（単語を $w$，関係マーカを $rel$）に対して，

$$\text{typicalness}(nd, rel, w, pos) = \begin{cases} \dfrac{\sum_{c\in C} oc(c,rel,w)}{\sum_{c'\in C\cup C'} oc(c',rel,w)} & \text{if } pos = \text{head} \\ \dfrac{\sum_{c\in C} oc(w,rel,c)}{\sum_{c'\in C\cup C'} oc(w,rel,c')} & \text{if } pos = \text{modifer} \end{cases}$$

ここで，$C$ は単語 $nd$ およびその下位（子孫）の単語からなる集合，$C'$ は $nd$ の兄弟およびその下位（子孫）の単語集合である．$oc$ は，2章で述べた係り受け関係の頻度ををを示している．**表3** に，typicalness > 0.5 の条件を満たす視点の例をあげる．

### 3.3 未知語配置アルゴリズムとその適用例

シソーラス上の各単語に対する視点集合が抽出された後，各未知語の配置は，以下の手順で行われる．

(1) 入力語に対して，シソーラス上のすべてのノード（単語）との間で類似度を計算する☆．
(2) 類似度が閾値を超えるノードをマークする．
(3) マークされたノードから，連結されたノードの集合（エリア）を求める．
(4) 各エリアを評価し，優先度を付与する．最も優先度の高いエリアを，入力単語の所属するエリアとする．

実際に，どのように未知語の位置が推定されるかを「戦闘機」という単語を例に見てみよう．まずこの単語とシソーラスに属するすべての単語との間で類似度を計算する．単語間の類似度 sim は，それぞれの単語の共起関係を用いて次のようにで与えられる．まず，単語 $w$ に対して，以下のようなベクトルのペア

$\langle X(w), Y(w)\rangle$
$X(w) = [x_{11}, x_{12}, \ldots, x_{1n}, x_{21}, x_{22}, \ldots, x_{2n}, \ldots, x_{m1}, x_{m2}, x_{mn}]$
$Y(w) = [y_{11}, y_{12}, \ldots, y_{1n}, y_{21}, y_{22}, \ldots, y_{2n}, \ldots, y_{m1}, y_{m2}, y_{mn}]$

を定義する．ただし，シソーラス中の全単語を $w_1, w_2, \ldots, w_n$，全関係マーカを $r_1, r_2, \ldots, r_m$ とし，$x_{ij} = oc(w, r_i, w_j)$，$y_{ij} = oc(w_j, r_i, w)$ とする．このとき，$w$ と $w'$ の類似度 $\text{sim}(w, w')$ を，

$$\max\left(\frac{X(w)\cdot X'(w')}{|X(w)||X'(w')|}, \frac{Y(w)\cdot Y'(w')}{|Y(w)||Y'(w')|}\right)$$

と定義する．上式中の $\cdot$ はベクトルの内積，$|X|$ はベクトル $X$ のノルムである．

☆ シソーラス上には，たとえば「水上・水中乗物」のように，実際にはテキストに出現しない単語がある．このような単語に対しては，それらの下位語の関係をランダムにとりだして，その単語の関係とする．とりだす（係り受け関係 $oc$ の）個数は，下位語それぞれが持つ $oc$ の個数の平均とする．

表3 抽出された視点の例
Table 3 Example of extracted viewpoints.

| 単語 | 視点 |
|---|---|
| 航空機 | (が, 飛ぶ,"head"), (が, 離陸する,"head"), (が, 着陸する,"head") |
| ロケット | (を, 発射する,"head") |
| 船 | (が, 入港する,"head"), (が, 沈没する,"head") |
| 陸上乗物 | (で, 輸送する,"head") |

表4 マークされた単語と一致した関係 (一部)
Table 4 Marked node with matched relationships.

| 順位 | 単語 | 関係 |
|---|---|---|
| 1 | 人 | 悪い～, ～を保護する |
| 2 | 株 | ～を買う, ～を持つ, ～を売る |
| 3 | 製品 | ～を買う, ～を持つ, ～を売る |
| 4 | 物 | ～を買う, ～を持つ, ～を売る |
| 5 | 金 | ～を買う, ～を持つ, ～を奪う |
| 6 | 住宅 | ～を買う, ～を持つ, ～を売る |
| 7 | 技術者 | ～を持つ, ～を派遣する |
| 8 | 企業 | ～を買う, ～を持つ, ～を保護する |
| 9 | 部品 | ～を買う, ～を売る, ～を輸出する |
| 10 | 設備 | ～を買う, ～を持つ, ～を奪う |
| 11 | 本 | ～を買う, ～を持つ, ～を奪う |
| 12 | 隊 | ～を派遣する, ～が危険, ～と衝突する |
| 13 | 航空機 | ～を買う. ～が飛ぶ, ～を売る |
| 14 | 兵器 | ～を買う, ～を持つ, ～を売る |

直観的には，ここでの類似度は，係り受け関係の類似性，言い換えれば，統語的な類似性を反映したものである．一般に，シソーラスは，主に意味的な類似性を重視しており，たとえば，ある分野では「製品」と「サービス」が同じカテゴリに分類されるといった知識を記述するのには適していない．

類似度 sim が，ある閾値を超えた場合，その単語はマークされ，マークノードリストに入れられる．表4に,「戦闘機」に対するマークノードリストに含まれるノードのうち，sim の値が大きいものを示す．

これらマークされた単語をエリアにまとめるのが次の処理である，これは，マークノードリスト中のノードのうち,(直接の) ISA 関係で連結された部分木を求めることで行われる．図4に，認識されたエリアを示す．

最後に，エリアの候補を評価して，最適のエリアを得る．それぞれのエリアは，次のような4つの基準で評価される．

**基準1 (C1):** エリアに含まれる単語との類似度の和．つまり未知語を $w$，エリア内の単語集合を $C$ とすると，$C1=\sum_{node\in C}\mathrm{sim}(w, node)$.

**基準2 (C2):** エリアの高さ．つまり，C2 = エリアが含むレベルの数．たとえば，図4の(a)のエリアでは，C2 = 2.

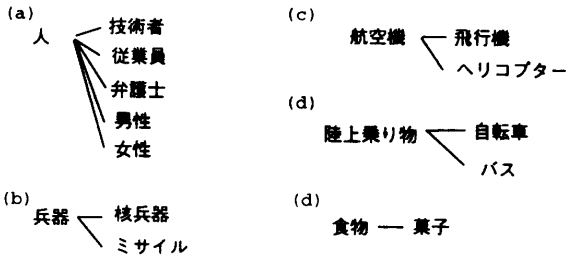


図4 "戦闘機" が所属するエリアの候補
Fig. 4 Candidate connections for "fighter".

**基準3 (C3):** シソーラスのトップノードからエリア内の最も上位の単語までの階層数．

**基準4 (C4):** 視点の数．たとえば，図4の(a)は，含まれる単語の数が一番多く，基準 C1 からは，一番優先性が高い．しかし，このエリアの中の単語，たとえば「人」と,「戦闘機」で一致した関係を見てみると,「悪い人/悪い戦闘機」「人が守る/戦闘機が守る」といった typicalness の値が小さな関係である．逆に，(c) 内の「飛行機」とは,「飛行機」の視点☆である「～が飛ぶ」という関係を共有しており，基準 C4 は，これらの関係を持つエリアに重みをかけるものである．具体的には，C4 = そのエリアに含まれる単語の視点が，入力単語の持つ単語間の関係に含まれている数，である．

ここで，基準1と基準2は，3.1節であげた規則(1)に，基準3は規則(2)に，基準4は規則(3)に，それぞれ相当するものである．

最終的な優先度 $P$(word) は，それらを重みつきで加算して求められる．

$$P(\mathrm{word}) = p_1C1 + p_2C2 + p_3C3 + p_4C4$$

ここで，$p_1$, $p_2$, $p_3$, $p_4$ は，重み係数である．本論文では，それぞれを，$p_1$=1, $p_2$=1, $p_3$=0.4, $p_4$=3 とした．これは，あらかじめ用意した2種類の訓練集合，(1) シソーラス中の単語そのもの，(2) 訓練用に用意した未知語の集合と，正解（その語がシソーラス上のどこに属すかを人手で付与したもの）に対して予備実験

☆ 定義より，視点は typicalness が高い

を行い，その結果を元に，未知語が正解の位置により多く配置されるように，パラメータを調整することで決定した．

上の式を用いて，最終的に，「戦闘機」が属す可能性の最も高いエリアは，(c) の「航空機」をトップノードとするものとなる．

## 4. 実　　験

前章までで述べた手法を用いて，ISAMAP 上に単語を配置する実験を行った．実験では，ISAMAP の“具体物”および“有意志体”を最上位とする単語約 2,000 語を用いた．**表 5** に結果の一部を示す．

実験から，視点の存在が，位置推定に大きな比重で寄与することが分かった．逆に，そのノードの視点が存在しない場合，単語の位置を精度良く推定することは困難である．従来の，統計手法を用いた単語のクラスタリング手法では，結果が人間の直観に合わないことが多いが，これは共起する単語が typicalness の値が低い値であると考えられる．

次に，推定したい単語の出現頻度と推定の精度の関係に関する実験を行った．コーパスから，約 20，50，100，200，300 の出現頻度を持つ単語を各々50 個ずつ用意し，推定の結果，優先度が最も高いエリアにその単語が属すと判定された割合を**図 5** に示した．なお，シソーラス上の単語に含まれる視点数の平均は 2.5 個であった．出現頻度が多くなるにつれて，精度は向上するが，頻度が非常に多いものに対しては，かえって精度が悪化してしまう．理由としては，頻度が増すにつれて，様々な関係が抽出されるので，類似度が閾値を超えるノードの数も増え，また，各ノードの視点となる関係も多く含まれてしまうという現象が生じる．その結果，候補となるエリアの数が増加し，優先度計算の精度が落ちるためだと思われる．実験では，出現頻度が 50～100 個の場合が，精度が一番高かった．

本論文で提案した手法は，未知語に対して，それが属すべきエリアを推定するものである．さらに，このエリアのどこにその単語が属すかを決定するためには，次のような情報が有効であると考えられる．エリアの階層が 2 のときには，下位の単語と姉妹関係である可能性が高い．また，シソーラスのサイズが比較的多い場合には，未知語が抽象度が非常に高いレベルに位置するとは考えにくい．また，コーパスから抽出できる表層的な情報も利用できる．たとえば，その語が並列関係にある語は，その単語のクラス決定の重要な手がかりとなる．実験で用いた“重油”の場合，これが属すべきエリアは (品物 (原材料 (燃料 (ガス，ガソリン)))) と推定されるが，「重油とガス」という係り受け関係の存在から，“重油”は，“ガス”と姉妹関係にある可能性が高いことが推定される．

**表 5** 実験結果
Table 5 Experimental result.

| 単語 | 位置 (エリア) |
|---|---|
| 裁判所 | (機関 (組合，会議，政党) |
| 大統領 | (人間 (男性，女性，弁護士，家族，社員，etc)) |
| オーストラリア | (国家 (日本，中国，ロシア)) |
| プレゼント | (品物 (食品，帽子，部品，etc)) |
| 自民党 | (機関 (組合，会議，政党，チーム，etc)) |
| 著者 | (人間 (男性，女性，弁護士，家族，社員，etc)) |
| セミナー | (施設 (学校，公共施設，駐車場，etc)) |
| 美術館 | (施設 (学校，公共施設，駐車場，etc)) |
| 妻 | (人間 (男性，女性，弁護士，家族，社員，etc)) |
| 重油 | (品物 (原材料 (燃料 (ガス，ガソリン)))) |

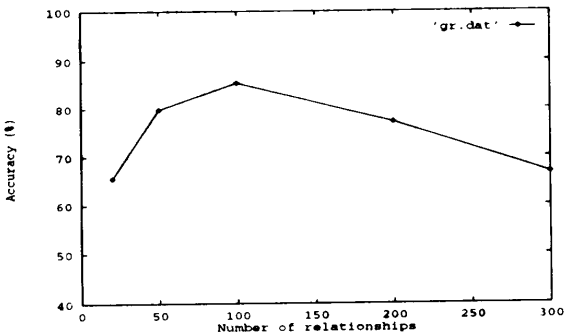


**図 5** 未知語の出現頻度と推定の精度の関係
Fig. 5 Relationship between the number of relationships and the accuracy of positioning.

## 5. コーパスに基づくシソーラスの構築

人間の内省に基づくシソーラスと，特定分野の知識を反映できるようなコーパスからの統計情報を用いることによって，コーパスに基づくシソーラスを構築することが可能である．このシソーラスは以下のものから構成される．

- コアシソーラス
- コーパスから抽出した統計情報（分野ごと）
- 単語間関係計算機構
- ユーザアクセスのためのアクセス関数

1 章で述べたように，シソーラスの工学的使用を考えた場合，シソーラスで扱えるすべての単語があらかじめ階層的に記述されている必要はなく，ユーザが利用するいくつかの関数を通じて，あたかも単語間の関係が記述されているかのようにシミュレートできればよい．本論文で提案しているように，一般分野のコアシソーラスとユーザが対象としている分野の統計情報

を用意し，関数群に対して値を計算できるような情報を提供できれば，コーパスの分野を適宜選択することで，分野によって変化し得る単語間の関係をユーザに提供することができる．ユーザがアクセスする関数には，次のようなものがある．

| | |
|---|---|
| position($w$) | 単語 $w$ の位置あるいはその単語が属す可能性の高いエリアを返す． |
| superordinate($w$) | 単語 $w$ の上位語（集合）を返す． |
| subordinate($w$) | 単語 $w$ の下位語（集合）を返す． |
| synonyms($w$) | 単語 $w$ の同義語集合を返す． |
| similarity($w1, w2$) | 単語 $w1, w2$ 間の類似度を返す． |

入力された単語がシソーラスに含まれていない場合でも，統計データを用いて，その位置を推定することにより，これらの関数に対して，解が与えられる．また，コーパスに基づくシソーラスのもう1つの利点は，シソーラスという主に単語間の上位下位関係を重視した知識源と，コーパスデータという共起関係を重視した知識源とをうまく統合できることである．たとえば，単語 $w1, w2$ 間の類似度 similarity($w1, w2$) は次のように定義されている．

$$\begin{aligned}&\text{similarity}(w1, w2)\\&\quad = \alpha \times \text{sim\_by\_thesaurus}(w1, w2)\\&\qquad + \beta \times \text{sim\_by\_cooc}(w1, w2)\end{aligned}$$

ここで，sim_by_thesaurus は，シソーラス上での類似性を計算する関数であり，sim_by_cooc は，共起関係から見た類似性を計算する関数である（これには，3.3節で定義した関数 sim が用いられる）．両者の関数の値は $[0, 1]$ で正規化される．$\alpha$，$\beta$，それぞれに対する重みであり，この重みをユーザが必要に応じて設定可能にすることで，単語間の類似性を動的に変化させることができる．たとえば，本シソーラスシステムでは，単語「戦闘機」に対する同義語（関数 synonyms によって得られる）は，パラメータに応じて，次のように変化する（上位3語を出力）．

$\alpha = 2.0, \beta = 1.0$ のとき，

**synonyms**(戦闘機) = (航空機, 飛行機, ヘリコプター)

$\alpha = 1.0, \beta = 2.0$ のとき，

**synonyms**(戦闘機) = (航空機, 兵器, 飛行機)

シソーラスの構造を重視する場合には，最も類似する語（この場合，航空機）の上位語や姉妹関係にある語が，同義語として上位にくる．しかし，統計情報を重視する場合，共起関係で類似する「兵器」のような単語が，同義語として選択される☆．

☆ もっとも，もともと「戦闘機」はシソーラス上にないため，最も類似する語は統計的な類似度を反映している．

## 6. おわりに

本論文では，コーパスに基づくシソーラスを構築するための基礎として，既存の中規模のシソーラスとコーパスを用いて，シソーラスを拡張する手法について述べた．既存のシソーラス作成者はある基準に従って，単語をシソーラス上に配置したはずであり．未知語のシソーラス上での位置を推定するのためには，シソーラスを構築するために用いられた基準を用いるということが考えられる．著者らは，この基準を「視点」と呼ぶ．本論文では，シソーラスの構造に沿って，視点と呼ばれる単語間の関係を用いることで，未知語が属す可能性が一番高いエリアを推定することが可能である．今後の研究としては，次のものがある．

また，未知語にも多義性がある場合がある．たとえば，「戦闘機」には，「兵器」としての意味があるが，これらの関係も含めて，未知語の妥当な意味を知ることは困難である．たとえば，テキストコーパスから，より多くの情報を抽出するなどの対策が必要となる．

**謝辞** ISAMAP の使用を快く許可していただいた東京工業大学の田中穂積教授に感謝いたします．

## 参考文献

1) Hearst, M.A. and Grefenstette, G.: Refining Automatically-discovered Lexical Relationships: Combining Weak Techniques for Stronger Results, *Proc. AAAI Workshop on Statistically-based NLP Techniques*, pp.64–72 (1992).
2) Li, X.: A WordNet-based Algorithm for Word Sense Disambiguation, *Proc. 3rd Annual Workshop on Very Large Corpora*, pp.1368–1374 (1995).
3) Miller, A., Beckwith, R., Fellbaum, C., Gros, D., Miller, K. and Tengi, R.: Five Papers on WordNet, Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University (1990).
4) Resnik, P.: Disambiguating Noun Grouping with Respect to WordNet Senses, *Proc. 3rd Annual Workshop on Very Large Corpora*, pp.54–68 (1995).
5) Yarowsky, D.: Word-sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, *Proc. COLING-92*, pp.454–460 (1992).
6) Yarowsky, D.: Unsurpervied Word Sense Disambiguation Rivaling Surperviced Methods, *Proc. ACL'95*, pp.189–196 (1995).
7) 田中，仁科：上位/下位関係シソーラス ISAMAP1

の作成 [1][2]，情報処理学会自然言語処理研究会，Vol.64, No.4, pp.25–44 (1987).

(平成8年1月29日受付)
(平成8年9月12日採録)

**浦本 直彦**（正会員）

昭和40年生．平成2年九州大学工学部総合理工学研究科情報システム学専攻修了．同年日本アイ・ビー・エム（株）入社．東京基礎研究所に配属．以来，機械翻訳システム，情報検索システムの研究開発に従事．平成3年度人工知能学会全国大会優秀論文賞授賞．言語処理学会，人工知能学会各会員．