社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180（15）
2007／7／24

ロボット型検索エンジンを用いた未知語の理解支援手法

後藤 和人† 渡部 広一‡ 河岡 司‡

†‡同志社大学大学院 工学研究科 知識工学専攻 〒610-0394 京都府京田辺市多々羅都谷 1-3

E-mail: †kgotou@indy.doshisha.ac.jp, ‡{hwatabe, tkawaoka}@mail.doshisha.ac.jp

あらまし 日常的な会話の中では，新語や固有名詞などのシソーラスに定義されていない語（未知語）が出現する．新語や固有名詞についての知識がなければ，会話を続けることは困難となる．そこで，それらの語に対する適切なシソーラスのノードを提示することによって，円滑な会話が可能となる．本稿では，シソーラスに定義されていない語を概念ベースと関連度計算方式，および，Web を用いて，各ノードとの関連性を評価し，最適なノードへ分類する手法を提案する．

キーワード シソーラス，概念ベース，関連度，未知語

Understanding Support Method of Unknown Words Using Search Engine

Kazuto GOTO† Hirokazu WATABE‡ and Tsukasa KAWAOKA‡

†‡Dept. of Knowledge Engineering & Computer Sciences, Doshisha University 1-3 Miyakodani, Tatara, Kyotanabe, Kyoto, 610-0394 Japan

E-mail: †kgotou@indy.doshisha.ac.jp, ‡{hwatabe, tkawaoka}@indy.doshisha.ac.jp

Abstract The words which are not defined in Thesaurus appear in the daily conversation, including new words and proper nouns. It is difficult to carry conversation with no knowledge about these words. The smooth conversation can materialize by presenting an appropriate node for these words. This paper proposes the technique for finding the best node for the word which is not defined in Thesaurus.

Keyword Thesaurus, Concept-base, Degree of Association, Unknown Words

1. はじめに

自然言語処理において，会話文中にシソーラスに定義されていない語（以下未知語とする）が含まれる場合，語の理解を行うことが困難である．そのため，未知語が大局的にどのような意味を持つのかを知る必要がある．未知語が所属すべきシソーラスのノードを提示することで，未知語の内容を簡明に表示することができると考える．

本稿では，シソーラスの体系的特徴を基に未知語が属すべき最適なノードへ分類する手法を提案する．

2. 使用技術

2.1. シソーラス[1]

シソーラスとは，一般名詞 2710 個の意味属性（ノード）が持つ上位下位，全体部分関係が木構造で示されたものである．

本稿では，未知語が属するノードを探す上で，ノード「設立」など未知語が所属するものとして不適切なノードを削除している．結果，使用するノード数は 370 個となっている（図 1）．また，ノードに登録されているリーフについては，手を加えることなく用いている．

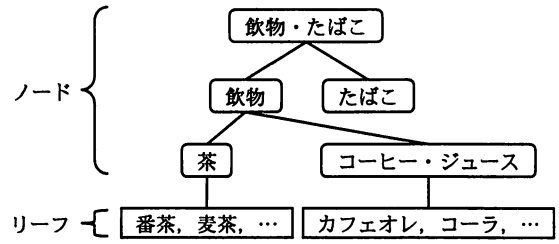

図 1 シソーラス（一部）

2.2. 概念ベース[2]

概念ベースとは複数の国語辞書や新聞などから機械的に構築した語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである．概念と属性のセットにはその重要性を表す重みが付与されている．概念ベースには，約 9 万語の概念が収録されており，1 つの概念に約 30 個の属性が存在する．

ある概念 A は属性 a_i とその重み w_i の対の集合として式（1）で表される．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_i,w_i),\cdots,(a_m,w_m)\} \quad (1)$$

任意の1次属性 a_i は，その概念ベース中の概念表記の集合に含まれている語で構成されている．したがって，1次属性は必ずある概念表記に一致するため，さらにその1次属性を抽出することができる．これを2次属性と呼ぶ．概念ベースにおいて，「概念」は n 次までの属性の連鎖集合により定義されている．

本稿では，概念ベースを Web からの未知語属性の獲得（3.1 参照），および，ノード属性の獲得（3.2 参照）に利用している．

2.3. 概念間の関連度計算方式[3]

関連度計算方式とは，概念と概念の間にある関連の強さを定量的に評価するものである．本節では関連度計算方式に用いる重み比率付き一致度と，一致度より算出される関連度の定義について述べる．

2.3.1. 重み比率付き一致度

任意の概念 A，B について，それぞれの1次属性を a_i，b_j とし，対応する重みを u_i，v_j とする．また，概念 A，B の属性数を L，M 個 $(L<M)$ とする．

$$A=\{(a_i,u_i)\mid i=1\sim L\},\quad B=\{(b_j,v_j)\mid j=1\sim M\}$$

このとき，概念 A，B の重み比率付き一致度 $MatchWR(A,B)$ を式（2），（3）で定義する．

$$MatchWR(A,B)=\sum_{a_i=b_j}\min(u_i,v_j) \tag{2}$$

$$\min(\alpha,\beta)=\begin{cases}\alpha & (\beta>\alpha)\\ \beta & (\alpha>\beta)\end{cases} \tag{3}$$

2.3.2. 関連度

概念 A と概念 B の関連度 $MR(A,B)$ を求めるアルゴリズムを以下に示す．

1. 概念 A と概念 B の属性を重み順に上位 t 個を抽出する．
2. 属性数の少ない方の概念を A とし，概念 A の属性を基準とする．$A=\{(a_1,u_1),(a_2,u_2)\cdots,(a_L,u_L)\}$
3. 属性同士が完全に一致する属性 $(a_i=b_j)$ は別扱いとする．すなわち，$\max(u_i,v_j)$ となる a_i または b_j の重みを $\max(u_i-v_j)-\min(u_i-v_j)$ とし，その重みを u_i' あるいは v_j' とする．
4. 完全に一致する属性を除いた概念 A' と概念 B' は，
$$A'=\{(a_1',u_1'),(a_2',u_2'),\ldots,(a_i',u_i'),\cdots,(a_L',u_L')\}$$
$$B'=\{(b_1',v_1'),(b_2',v_2'),\ldots,(b_j',v_j'),\cdots,(b_M',v_M')\}$$
のように表せる．ここで，概念 A' と概念 B' には完全に一致する属性が存在しないので，類似する属性について一致度を計算する．1次属性の一致度により全体の一致度の和が最大となるように対応を決め，定まった各属性の一致度とそれぞれの属性の重みを用いて，関連度を求める．
5. 完全に一致する属性が α 個あったとすると，概念 A と概念 B の意味関連度を式（4）で定義する．

$$MR(A,B)=\sum_{i=1}^{\alpha}\{\max(u_i-v_j)-\min(u_i-v_j)\}+\sum_{i=1}^{t+\alpha}MatchWR(a_i',b_j')\times\frac{u_i'+v_j'}{2}\times\frac{\min(u_i',v_j')}{\max(u_i',v_j')} \tag{4}$$

上記の式より，関連度は概念間の関連の強さを 0 と 1 の間の実数値で表す．表 1に関連度計算の例を示す．

本稿では，未知語属性とシソーラスのノード属性との関連の深さを判断するために関連度計算を用いる．

表 1　関連度計算の例

概念 A	概念 B	関連度
自動車	車	0.912
自動車	学校	0.001

2.4. 概念間の関連度計算方式

本節では，本稿が提案する手法で利用した重み付け手法である tf・idf[4] と SWeb-idf[4] について述べる．

2.4.1. tf・idf

tf・idf による重み付けとは，語の頻度と網羅性に基づいた重み付け手法である．文書 d における索引語 t の重み $w(t,d)$ は以下の式（5）によって得られる．

$$w(t,d)=tf(t,d)\cdot idf(t) \tag{5}$$

$tf(t,d)$ は文書 d における索引語 t の出現頻度である．また，$idf(t)$ は検索対象文書数 N と索引語 t が出現する文書の数 $df(t)$ によって決まり，式（6）によって定義される．

$$idf(t)=\log\frac{N}{df(t)}+1 \tag{6}$$

本稿では，この重み付け手法をシソーラスのノード属性の重み付け（3.2 参照）に用いている．

2.4.2. SWeb-idf

SWeb-idf（Statics Web Inverse Document Frequency）とは，Web 上の語の idf を統計的に調べた idf 値である．まず，無差別に選んだ固有名詞 1000 語を作成する．この作成した 1000 語に対して個々に Google[5] で検索を行い，1 語につき 10 キャッシュを取得する．よって，得られたキャッシュ数は 10000 キャッシュとなる．この 10000 キャッシュを Web の全文書空間と見なし，その中での語の出現割合を求める SWeb-idf は，式（7）

で求められる．これらにより得られた語とその idf 値をデータベースに登録した．なお $df(t)$ 項は，全文書空間(10000 キャッシュ)に出現する概念 t の頻度である．

$$SWeb-idf(t)=\log\frac{N}{df(t)} \quad (N=10000) \qquad (7)$$

本稿では，この重み付け手法を Web より獲得する未知語属性の重み付け（3.1 参照）に用いている．

3. 未知語のシソーラスノードへのマッピング

本稿が提案する手法では，未知語を入力した後に，未知語とノードの属性を概念ベースに存在する語から獲得する．そして，得られた属性を用いてシソーラスのノード決定を行う．

処理の流れとしては，まず，未知語を入力した後に，未知語とノードの性質を比較する(関連度計算を行う)ために未知語属性とノード属性を獲得する．次に，獲得した属性群を用いて関連度計算を行い，所属ノード候補を絞り込む．さらに，シソーラスが持つ情報を利用（所属ノード決定手法）して，未知語が所属すべきノードを決定する．図 2に未知語をシソーラスのノードへマッピングする流れを示す．

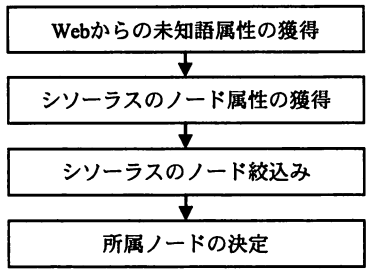

図 2　ノード決定の流れ

3.1. Web からの未知語属性の獲得

3.1.1. 未知語の概念化

未知語を入力して Google で検索を行い，検索結果ページを取得する．不要な情報を取り除いた文書群の形態素解析[6]を行い，自立語を抽出する．最後に概念ベースに存在する語を未知語の属性とする．そして，得られた属性の頻度に SWeb-idf (2.4.2 参照) の値を掛け合わせたものを属性の重みとし，重み順に並び替える．なお，SWeb-idf のデータベースに存在しない属性については SWeb-idf 値の最大値を掛け合わせている．

表 2　未知語「クイニーアマン」の属性（一部）

属性	重み
パン	382.88
バター	249.17
菓子	116.73

3.1.2. 未知語属性の拡張

未知語属性の拡張として，未知語の属性を 2 次属性まで展開して獲得する．これは 3.1.1 で説明した手法により獲得した未知語の属性（1 次属性）をキーワードとして再び Google で検索を行い，属性（2 次属性）の獲得を行うものである．表 3に未知語「クイニーアマン」を 2 次属性まで展開した様子を示す．

表 3　未知語「クイニーアマン」の属性
（2 次属性まで展開）

	属性 重み	属性 重み	属性 重み
1 次属性	パン 382.88	バター 249.17	菓子 249.17
2 次属性	パン 1457.08	バター 2062.55	菓子 1805.81
	酵母 107.04	発酵 149.49	和菓子 235.57
	レシピ 86.68	サンド 128.25	ケーキ 204.19

得られた 2 次属性をそのまま加えると，2 次属性の影響力が大きくなるという問題に対処するため，1 次属性の重みの大きさを考慮している．具体的には，2 次属性の重みに 1 次属性の重みの比率を掛け合わせている．例えば，表 3における 1 次属性「パン」の重みの比率は以下のように計算される．

382.88/748.78≒0.51

(「パン」の重み/1 次属性の重みの合計)

1 次属性「パン」の属性である「パン」，「酵母」，「レシピ」の重みは以下のように計算される．

- 「パン」: 1457.08*0.51=743.11
- 「酵母」: 107.04*0.51=54.59
- 「レシピ」: 86.68*0.51=44.21

最終的に得られる未知語「クイニーアマン」の属性を表 4に示す．

表 4　2 次属性まで展開して得られた未知語
「クイニーアマン」の属性

属性	重み	属性	重み
パン	1125.99	レシピ	44.21
バター	929.81	サンド	42.32
菓子	405.66	和菓子	37.69
酵母	54.59	ケーキ	32.67
発酵	49.48		

3.2. シソーラスのノード属性の獲得[7]

各ノードに属する全てのリーフに対して概念ベース参照を行い，リーフを概念とする語の 1 次属性とその重みを取得する．そして，これらを足し合わせたも

のを属性集合として取得する．この作業を全てのノードに対して行い，シソーラスのノード属性を取得する．次に，取得したシソーラスの全ノード属性内で，tf・idf（2.4.1 参照）を利用して各属性の重みを求める．これをシソーラスのノード属性とする．表 5に例として「時計」のノード属性を示す．

表 5 ノード「時計」の属性（一部）

ノード属性	重み
懐中時計	4733.49
掛時計	3476.39
置時計	2791.44

3.3. シソーラスのノード絞込み

3.1 で説明した手法を用いて属性を獲得した未知語と 3.2 で説明した手法を用いて属性を獲得したノード属性に対して関連度計算を行う．なお，関連度の閾値を 0.0 から 0.05 まで 0.001 刻みで実験を行った結果，最も高い精度を得られた 0.02 以上の関連度を持つノードを所属ノード候補とする．

3.4. 所属ノード属性の決定

3.4.1. ノード動詞

シソーラスは単語を体系的に配置しており，「同一のノードに属するリーフは助詞を伴う動詞の係り受けに同様の語を取る」という関係が存在する．ノード動詞とはこの関係を利用して，ノードに設定したキーワードのことであり，ノード決定の補助に利用する．

例えば，未知語が「マイルドセブン」，所属ノード候補が「たばこ」である場合，「マイルドセブンを吸う」というキーワードの検索を Google で行ったときの HIT 数を求める．表 6にノード動詞の例を示す．

表 6 ノード動詞（一部）

ノード	ノード動詞
飲物	を飲む
菓子	を食べる
カメラ	で撮影
たばこ	を吸う
歌手	を歌う

3.4.2. 共起ヒット

関係のある 2 語はある文書に共に出現すると考えられる．そこで，未知語とノード名の And 検索による HIT 数を調べてノード決定の補助を行う．

例えば，未知語が「マイルドセブン」，所属ノード候補が「たばこ」である場合，「マイルドセブン」と「たばこ」で And 検索を Google で行ったときの HIT 数を求める．

3.4.3. 所属ノードの決定手法

未知語の所属ノードの決定を以下の 4 つの手法で行った（表 7）．ノード得点の計算式は以下の式(8)，(9)，(10)，(11) に示したものであり，所属ノード候補 $node_i$ の中でノード得点 $NodeValue$ が最も高いノードを所属ノードとする．MR が未知語と $node_i$ の関連度，$VerbHit(node_i)$ は未知語にノード動詞を連結したキーワードの検索を Google で行ったときの HIT 数，$CoincidenceHit(node_i)$ は未知語とノード名の And 検索を Google で行ったときの HIT 数を表す．

表 7 所属ノードの決定手法一覧

手法	所属ノード決定手法	計算式
①	獲得した未知語属性と ノード属性の関連度計算	式（8）
②	①＋ノード動詞	式（9）
③	①＋共起ヒット	式（10）
④	①＋ノード動詞＋共起ヒット	式（11）

$$NodeValue(node_i) = MR(node_i) \quad (8)$$

$$NodeValue(node_i) = MR(node_i) \cdot \log(VerbHit(node_i)) \quad (9)$$

$$NodeValue(node_i) = MR(node_i) \cdot \log(CoincidenceHit(node_i)) \quad (10)$$

$$NodeValue(node_i) = MR(node_i) \cdot \log(VerbHit(node_i)) \cdot \log(CoincidenceHit(node_i)) \quad (11)$$

3.5. 評価と考察

本稿で提案している未知語のノードの決定の評価を行うために，200 個の未知語とその語が所属するノードを対にしたテストセットを用いる．評価に使用したテストセットの一部を表 8に示す．

表 8 テストセット（一部）

未知語	所属ノード
クイニーアマン	パン
マイルドセブン	たばこ
新島襄	教師
FinePix	カメラ
Gショック	時計

テストセットを使用し，未知語の所属ノードの決定を行う．テストセットの各未知語の入力に対して，システムが返答した結果がシソーラスのノードの中で最も適切な答えが得られた未知語を正解，得られなかった未知語を不正解として精度を算出する．また，未知語 1 語あたりの平均処理時間を算出している．評価結果を図 3，図 4に示す．評価は 3.4.3 で述べた手法を用いている．

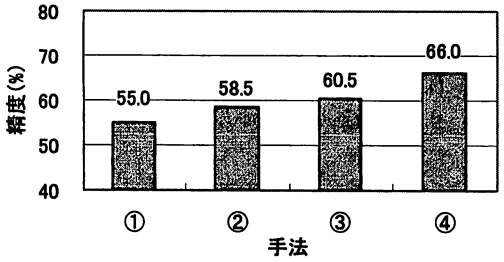

図 3　手法ごとの精度

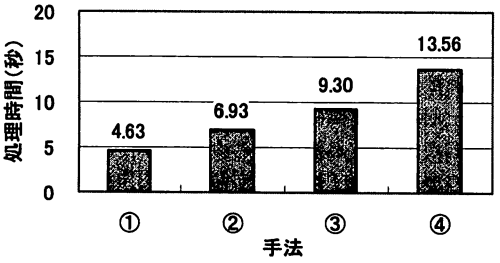

図 4　手法ごとの処理時間

図 3より，関連度計算のみでノードの決定を行う手法(手法①)では精度が 55.0%であったことに対して，ノード動詞や共起ヒットを利用する手法（手法④）では 66.0%の精度を得ることができた．

次に，3.1.2 で説明した未知語の属性を 2 次属性まで展開して獲得したときの評価を図 5，図 6に示す．なお図 5，図 6における 2 次属性を獲得する 1 次属性の個数は，獲得した未知語の 1 次属性の上位何件から 2 次属性を獲得したかを表す．また，所属ノード決定手法としては最も高い精度が得られたノード動詞と共起ヒットを利用する手法（手法④）を選択している．

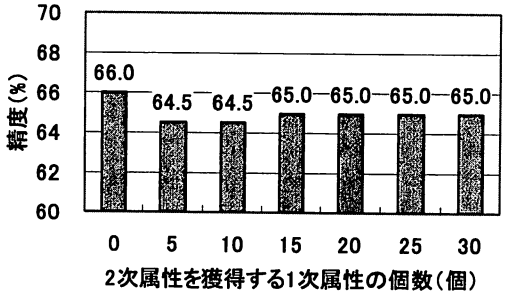

図 5　未知語の属性を
2 次属性まで展開して獲得したときの精度

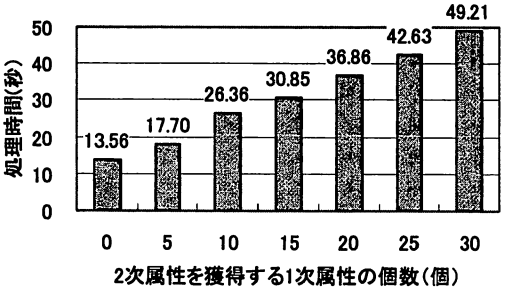

図 6　未知語の属性を
2 次属性まで展開して獲得したときの処理時間

また図 5，図 6より，未知語の属性を 2 次属性まで展開して獲得しても必要な処理時間が増えるだけで，精度の向上は見込めないことが分かった．

4. 複数のノードの提示

図 3より，手法④では 66.0%の精度が得られていることが分かる．この結果をノード得点の上位 3 件のノードまで広げたときの評価を図 7示す．例えば，未知語「新島襄」は所属ノードが「教師」であり，シソーラスマッピングを行うと，第 1 候補ノードが「教育」，第 2 候補ノードが「学校」，第 3 候補ノードが「教師」と出力される．したがって，第 3 候補ノードである「教師」が所属ノードであることが分かる．

図 7より，ノード得点上位 3 件の候補ノードを提示した場合には，76.5%の精度が得られていることが分かる．本節では，ノード得点上位 3 件の候補ノードから所属ノードを抽出する手法について述べる．

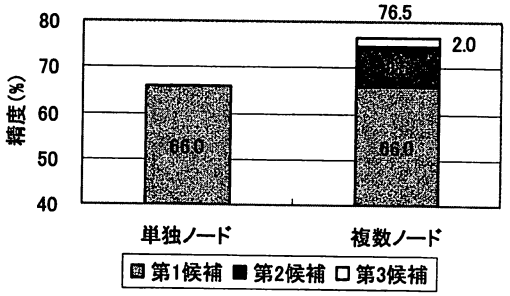

図 7　複数のノードを提示したときの評価

4.1. 親ノードとの関連度

2.1 で述べたようにシソーラスは，一般名詞の意味属性（ノード）の上位下位，全体部分関係が木構造で示されている．例えば，「茶」ノードの親ノードは「飲物」であるように，あるノードの親ノードは，そのノ

ードの上位語であり，より広い意味の言葉となる．したがって，親ノードはそのノードを大まかに説明したものといえる．そこで，未知語と所属ノードの親ノードにも関連があると考え，未知語と抽出したノード得点上位3件の候補ノードの親ノードとの関連度を比較し，最も関連度が高いノードを所属ノードとする．例として，表 9に未知語「新島襄」とノード得点上位 3件の候補ノードの親ノードとの関連度を示す．

表 9　未知語「新島襄」と
候補ノードの親ノードとの関連度

候補ノード	候補ノードの親ノード	関連度
教育	指導等	0.043
学校	公共機関	0.000
教師	教師・学生	0.065

4.2. 評価と考察

3.5 で述べた評価手法を用いて，本節で提案したノード得点上位3件の候補ノードから所属ノードを抽出する手法の評価を行う．なお，4.1 で述べた親ノードとの関連度を用いる手法を手法⑤としている．親ノードとの関連度を用いる手法を用いたときの評価を図 8，図 9に示す．

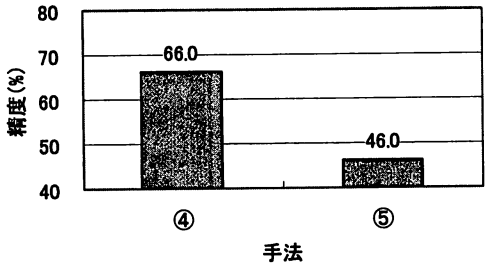

図 8　候補ノードから
所属ノードを抽出したときの精度

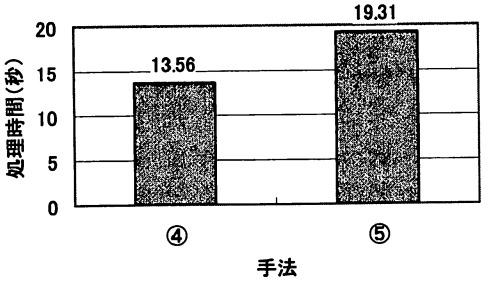

図 9　候補ノードから
所属ノードを抽出したときの処理時間

図 8より，親ノードとの関連度を用いる手法（手法⑤）はノード動詞と共起ヒットを利用する手法（手法④）と比較して精度が低下していることが分かる．精度が低下した原因としては，以下の理由が考えられる．

親ノードとの関連度を求める手法では，親ノードは親ノード自身が持つ子ノードそれぞれに関係する属性を全て持っている．そのため，未知語との関連性が薄れた結果，全体的に低い値が算出され，正確に候補ノードの比較を行うことができなかったと考えられる．

5. おわりに

本稿では，シソーラスに定義されてない語（未知語）が大局的に見てどういうものであるかを，シソーラスのノードにマッピングして提示する手法を提案した．これにより，文中に未知語が含まれる場合でも，未知語をノードに置き換えることで円滑な自然言語処理を行うことができると考えられる．

今後の課題としては，未知語の属性獲得手法の改良や未知語の属性から所属ノードを決定することでさらなるシステムの精度向上を目指したいと考えている．具体的には，未知語と関連が深い重み上位の属性のみを用いて属性の拡張を行うことや未知語の属性が多く所属するノードを所属ノードと導くことができるようにシステムを改良したいと考えている．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

文　献

[1] NTT コミュニケーション科学研究所監修，“日本語語彙体系”，池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦（編），岩波書店，1997.

[2] 小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構成法－語間の論理関係を用いた属性拡張”，自然言語処理，Vol.11，no.3，pp.21-38，Jul.2004.

[3] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，no.1，pp.53-74，Jan.2006.

[4] 辻泰希，渡部広一，河岡司，“www を用いた概念ベースにない新概念およびその属性獲得手法”，第 18 回人工知能学会全国大会論文集，Vol.2D1-01，Jun.2004.

[5] Google：http://www.google.co.jp

[6] 奈良先端科学技術大学院大学：
http://chasen.naist.jp/hiki/ChaSen/

[7] 伊藤俊介，渡部広一，河岡司，“情報検索における未知語理解支援方式～未知語のシソーラスノードへの分類～”，情報処理学会自然言語処理研究会資料，Vol.2004-NL-159，pp.61-66，Jan.2004.