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

2007－NL－180 (14)
2007／7／24

概念の揺らぎを考慮した概念間の関連度計算方式

奥村 紀之† 渡部 広一‡ 河岡 司‡

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

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

あらまし 本論文では，語と語の関連の強さを定量的に扱うための関連度計算方式について論じている．語と語の関連の強さを定量化するための代表的な手法として，ベクトル空間モデルがある．この手法では，語を定義するために属性を意味的，あるいは統計的に圧縮することにより基本ベクトルを定義し，ベクトル間の内積を用いて，概念間の類似度を定量化している．一方，本論文で対象としている関連度計算方式では，演算の基となる属性を圧縮することなく，関連する語の集合として概念を定義している．すなわち，本論文で対象としている概念ベースには概念の多義性が含まれている．そのため，語と語の関連の強さは，対象とする語間に共通している属性，類似している属性を動的に抽出し対応付けることにより算出している．したがって，演算対象となる語に多義性があった場合においても，柔軟に属性と属性の対応を決定し，一元的に定量化することが可能となる．

キーワード 概念ベース，関連度計算方式，多義性

A Calculation Method of Degree of Association between Concepts Considering Fluctuation of Concepts

Noriyuki OKUMURA† Hirokazu WATABE‡ and Tsukasa KAWAOKA‡

†‡Dept. Of Knowledge Engineering & Computer Sciences Graduate School of Engineering,

Doshisha University Kyo-Tanabe, Kyoto, 610-0394, Japan

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

Abstract In this paper, the calculation method of Degree of Association to treat depth of the relation between concepts quantitatively is discussed. The vector space model is a typical technique to quantify strength of the similarity between concepts. In this technique, the degree of similarity between concepts is quantified by defining a basic vector by meaning or statistically compressing the attribute to define the word and using the product in between vectors. On the other hand, the concept is defined as sets of concepts that relate without compressing the attribute that becomes the radical of the operation in the Degree of Association method targeted with this thesis. That is, the ambiguity of the concept is contained in the concept base targeted with this thesis. Therefore, depth of the relation between concepts is calculated by dynamically extracting and associating attributes common between targeted concepts and similar attributes. Therefore, flexibly deciding the correspondence of the attributes the attribute even if there is a ambiguity in the concept to be operated, and quantifying it uniformly become possible.

Keyword Concept-base, Calculation Method of Degree of Association, Ambiguity

1. はじめに

人間とコンピュータの円滑なコミュニケーションを実現するためには，コンピュータに人間の発する語に対する知識を付与する必要がある．本論文では，コンピュータに付与する知識として，国語辞書や新聞記事などから作成された概念ベースを用いている．

概念ベースにはおよそ8万7千の語が定義されている．しかし，概念ベース[1][4]を構成する各概念には，多義性を持つものも多く存在する．そのため，概念「トラック」が入力された場合，自動車の意味合いでのトラックであるか，陸上競技の意味合いでのトラックであるかを定量的に評価することが難しい．

概念ベースの多義性を解消するための手法としては，大きく二つの手法が考えられる．一つは概念ベースに定義される概念を，一義の概念によって定義する手法である．すなわち，概念を構成する属性を一義の概念に置換する手法である．もう一つは，概念ベースに定義される概念の多義性に依らず，関連度計算方式により選択的に属性を使用し，多義性を含む語と語の関連の強さを定量化する手法である．関連度計算方式[3]によって多義性を解消するためには，関連度計算の際に利用される属性を決定し，概念の意味を推定する

必要がある．

本論文では，概念を定義する属性を一義化する手法，関連度計算方式によって属性を選択的に利用し，多義性を解消する手法を比較検討する．

2. 概念ベース

2.1. 概念ベースの定義と構成

本論分で対象とする概念ベースは，概念を基底ベクトルの集合として定義するのではなく，概念(A)を，その概念を意味的に特徴づける概念や，その概念から容易に連想できる語(概念)の集合として定義する．概念を特徴づける概念を，概念に対する属性$(a_1,a_2,\cdots,a_n)$と呼ぶ．また，各属性にはそれら属性がどの程度概念を特徴づけているかを示す重み$(w_1,w_2,\cdots,w_n)$を付与している(式１)．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (1)$$

各概念を特徴づける属性$(a_1,a_2,\cdots,a_n)$もまた，概念ベースにおいて概念として定義されているため，$(a_1,a_2,\cdots,a_n)$もまた概念である．従って，属性の属性を二次属性，二次属性の属性を三次属性，といったように，無限次元の属性連鎖集合によって概念を定義している．

2.2. 概念ベースの構成法

現在，概念ベースは国語辞書より機械的に構築したもの（初期概念ベース）に加え，新聞記事から概念と属性を取得し概念ベースに登録することにより，約 8 万 7 千語の概念で構成されている．また，各概念に対し平均 30 属性が付与され，属性に付与される重みは，各概念について，その総和が 1.0 となるように正規化されている．

各概念表記に付与されている属性は，概念ベースに概念表記として登録されている語で構成されるため，各属性を一つの概念表記としてみなした場合，さらにそれを表す属性を導くことができる．すなわち，概念ベースにより，概念を n 次の属性連鎖集合として定義する．n 次の属性集合を n 次属性と呼ぶ．なお，次節に述べる関連度計算方式では，属性連鎖により概念表記を 2 次属性まで展開し演算を行う．以下「1 次属性」を単に「属性」とも表記する．また，語の意味(概念)とは，n 次の属性連鎖集合によって定義された属性空間における重み付き属性のことを示し，概念表記とは，語の意味を識別するラベル(例えば車や飛行機など)を示す．以下，概念ベースを CB と表記する場合がある．

2.3. 概念ベースの問題点

概念ベースにはおよそ 8 万 7 千語の概念が定義されているが，語の多義性について考慮されていないという問題点がある．本論文で対象とする多義性とは，概念の多義性と属性の多義性である．

概念の多義性とは，「バス」という語が入力された際に，「乗り合い自動車」の意味で使用されているのか，「男声」という意味で使用されているのかが識別できないという問題である．

また，属性の多義性とは，概念の多義性が解消されていないために起こる問題であり，例えば，概念：タクシーが「自動車，運賃，バス，…」といった属性によって定義されていた場合，属性：バスが「乗り合い自動車」の意味合いで付与されていると考えるのが自然であるが，概念の多義性によって，どの意味のバスであるのかを特定するのが困難であるという問題である．

本論文では，これらの多義性についてその解消手法について検討する．

3. 関連度計算方式

3.1. 意味関連度計算方式

意味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意味関連度の定義について述べる．

3.2. 一致度

任意の概念 A,B について，それぞれの概念に対する一次属性を 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 個，M 個($L<M$)とする．

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,w_L)\} \quad (3)$$
$$B=\{(b_1,v_1),(b_2,v_2),\cdots,(b_M,v_M)\}$$

このとき，概念 A，B の一致度 $MatchWR(A,B)$を以下の式で定義する．

$$MatchWR(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha(\alpha<\beta)\\ \beta(\beta\le\alpha)\end{cases} \quad (4)$$

概念 A，B の属性 a_i，b_j に対し，$a_i=b_j$（概念 A，B に共通する属性がある）となる属性があった場合，共通する属性の重みの共通部分，つまり，重みの小さい分だけ有効に一致すると考え，このように一致度を定義する．定義から明らかな様に両概念の属性と重みが完全に一致する場合に一致度は 1.0 となる．

3.3. 意味関連度

概念 A と概念 B の関連度 $MR(A,B)$は，

(1) 概念 A と概念 B の属性を重み順に上位 t 個抽出する．

(2) 属性の少ない方の概念を A とし，概念 A の属性を

基準とする．

$$A=\{(a_1,u_1),(a_2,u_2),\cdots,(a_L,w_L)\}$$

(3) 概念 B の属性を，概念 A の各属性との重み比率付一致度 $MatchWR(a_i,b_{x_i})$ の和が最大になるように並び替える．

$$B=\{(b_{x_1},v_{x_1}),(b_{x_2},v_{x_2}),\cdots,(b_{x_L},v_{x_L})\}$$

このとき，属性同士が完全に一致する場合（$a_n=b_{x_n}$）は別扱いとする．

これは，約 8 万 7 千語の概念によって定義される概念ベースにおいて，属性同士が完全に一致することは非常に稀であり，完全に一致した場合のみ，別扱いとすることで関連性を大きく評価するという意図に基づく．

具体的には，u_n と v_{x_n} を比較し，重みの小さい方にそろえ，重みの大きい方と一致分の差分を他の属性と再び対応させることで，より大きく関連性を評価することが可能となる．

このように一次属性の一致度により対応を決め，対応にあふれた概念 B の属性(b_{x_j})は無視する．

(4) 完全一致する属性が α 個あったとすると，概念 A と概念 B の意味関連度を以下に示す式で定義する．

$$MR(A,B)=\sum_{i=1}^{t+\alpha} MatchWR(a_i,b_{x_i})\times\frac{u_i+v_{x_i}}{2}\times\frac{\min(u_i,v_{x_i})}{\max(u_i,v_{x_i})}$$

このように関連度を定義しているため，関連度は如何に正しい属性が付与され，適切な重みが与えられているかに大きく依存する値となる．

3.4. 意味関連度の問題点

関連度計算方式では，語の多義性に関わらず，共通する属性，一致度の高い属性を使用し，演算を行う．そのため，本来対応すべきでない属性同士が対応し，関連度計算を行ってしまう問題点がある．

4. 概念ベースの意味分割

概念ベースの多義性を解消する手法として，概念構造情報を用いた意味分割手法が提案されている[7]．この手法では，辞書の記載情報から概念の意味が複数存在するものを取得し，一つの概念に対して複数の意味が割り当てられることなく，一つの概念には一つの意味を持たせることにより，多義性を解消している．

4.1. 概念構造情報

概念構造情報[2]とは，国語辞書から作成された概念ごとの意味や論理関係を示す知識である．本論分では，概念ベースの多義性解消手法として，概念の一義化，並びに関連度計算方式による概念の意味推定を行うが，すべて概念構造情報に基づいて処理を行うものとする．

表 1　概念構造情報の例　“星”（見出し語）

意味番号	関連語（関連語，関係型）
1	（星雲，類義）（天体，上位）・・・
2	（階級，不明）（記憶，不明）・・・
3	（点，上位）（斑点，上位）・・・
4	（眼球，不明）（白い，不明）・・・
5	（成績，上位）（白星，上位）・・・
6	（目当て，同義）（目ぼし，同義）・・・
7	（犯人，同義）・・・
8	（運勢，同義）（生まれる，不明）・・・
9	（移る，不明）（形，不明）・・・
１０	（スター，同義）（花形，同義）・・・

4.2. 概念の意味分割

表 1 に示した概念構造情報を基に，複数の意味を持つ概念(多義性を持つ概念)の意味分割を行う．概念構造情報において概念：星は，10 の意味によって構成されていることが分かる．多義性を考慮しない概念ベースでは，星を定義する ID が一つであったため，星の属性の中に，天体の意味の属性，犯人の意味の属性などが混在していた．これらの問題を解決するために，概念：星を定義するための ID を複数準備することにより，星 1～星 10 までの概念：星を再定義している．表 2 に星の中でも犯人に関する属性のみを選別した概念を示す．

表 2　概念：星の「犯人」での意味属性

1 次属性					
犯人	容疑者	犯罪	警察	逮捕	
犯す	被疑者	罪	国民	現行犯	2 次属性
罪	犯罪	犯す	生命	自由	
現行犯	起訴	重犯	都道府県	令状	
犯罪	法律	刑罰	財産	抑留	
当人	取調	刑法	国家	警察	

このように分割された概念ベースを多義分割 CB と定義する．

5. 概念の揺らぎを考慮した関連度計算方式

5.1. 多義語知識ベース

4.1 で示した概念構造情報を基に，多義語知識ベースを作成する．多義語知識ベースとは，関連度計算を行う際に，多義語を一義の概念に置換するためのテーブルである．表 3 に多義語知識ベースの例を示す．表 3 において，実概念とは，多義語が置換されるべき概念の中で，概念ベースに定義されている一義の概念(代表語)のことを示す．また，その他の分類には，代表的な意味分類以外のすべての意味を保持させている．

表 3　多義語知識ベース(概念：バス)

意味番号	代表語	
1	乗合自動車	実概念
2	浴室	実概念
3	コントラバス	実概念
4	バス	その他

5.2. 概念の揺らぎを考慮した関連度計算方式[8]

多義語の意味を推定するには，手がかりとなる語が必要である．例えば，表 3 における多義語：バスの意味を推定するための手がかりとして，「タクシー」という一義の概念を与えた場合，バスの意味分類の中で「乗合自動車」が最も関連が強いと推定することが出来る．そこで多義語と手がかり概念の二つの概念を与えた時の多義性解消の手順を以下に示す．

【基準が多義で対象が一義の場合】

- 基準概念の全ての代表語と対象概念で関連度計算を行い，基準概念を最高関連度の代表語に置換する．

【基準・対象ともに多義の場合】

- 基準概念の全ての代表語と，対象概念の全ての代表語で，関連度計算を行い，それぞれの概念を最高関連度の代表語に置換する．

このように，2 つの概念が与えられた時，多義の概念を代表語に置換し関連度計算を行う．拡張した関連度計算方式を，概念の揺らぎを考慮した関連度計算方式(多義性解消関連度計算方式)と呼ぶ．また，関連度計算を行う対象の 2 語が双方とも一義である場合には通常の関連度計算を行う．

また，多義語知識ベースを用いて，手がかりの語から意味分類を推定した際に「その他」の分類に該当した場合(置換すべき代表語が無い場合)においては，入力された多義語を定義する属性の中から，通常の関連度計算方式によって手がかりとなった語の意味に近いと推定される属性を動的に抽出し，概念を再定義する．図 1 にバスに付与された手がかりが男声であった場合の属性抽出の例を示す．

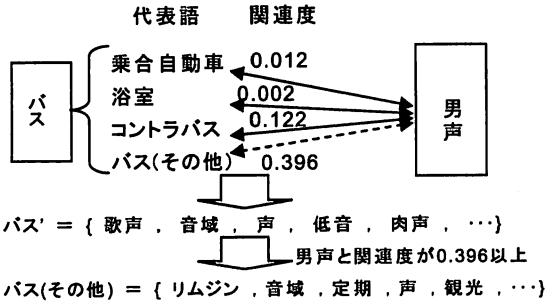

図 1　バスの属性から男声と関連の強い属性を抽出

本節以降，意味関連度計算方式を MR と表記したことに対し，多義性解消関連度計算方式を CR と表記する．

6. 評価手法

本節では，4.で構築した多義分割 CB，並びに，5.で述べた多義性解消関連度計算方式を評価する手法について述べる．

6.1. 評価用データ

評価尺度として，4 組 1 セットとなる概念表記群を準備する．これを(*X-A,B,C*)評価用データ（表 4）と呼び，任意の基準概念表記 *X* に対し，同義・類義等最も関連が深いと考えられる概念表記 *A*，概念表記 *A* ほどではないが関連があると思われる概念表記 *B*，そして，まったく無関係である概念表記 *C* によって構成する．このような評価用データを人手によって作成し，1980 組のデータを用意した．

評価用データの各セットは，複数の人間によって作成され，そのデータを作成者以外の 3 人の人間に判定させ，全員一致で正しいとされたもののみを利用する．また，評価用データに採用する概念表記は，概念ベースで定義されているものとする．

表 4　(*X-A,B,C*)評価用データ

X	*A*	*B*	*C*
音楽	楽曲	音	電車
学生	生徒	学業	林檎
・・・	・・・	・・・	・・・

また，作成した 1980 組の評価用データの中には，基準となる概念 *X* が多義語である評価セットが含まれている．表 5 に基準概念 *X* が多義である評価セットの一部を示す．また，無関連概念 *C* に関しては，*X* に対して *A* の観点で見た場合無関連であるが，別の観点で見た場合，高関連となる概念を選定している．

表 5　X が多義である評価用データ

X	*A*	*B*	*C*
バス	タクシー	運転手	風呂
星	天体	宇宙	警察
・・・	・・・	・・・	・・・

6.2. 評価手法

6.1.で述べた(*X-A,B,C*)評価用データを用いて，多義分割 CB と多義性解消関連度計算方式の性能を評価する手法について述べる．

(*X-A,B,C*)評価用データは，人間の感覚において，基準概念に対して(*X* と *A*)＞(*X* と *B*)＞(*X* と *C*)の順に関連

が強いと判定されている．したがって，概念ベースを用いた関連度計算においても，$MR(X,A) > MR(X,B) > MR(X,C)$と判定されたとき，正しく関連の強さを評価できていると言える．1980組の$(X\text{-}A,B,C)$評価用データのうち，正しく判定された$(X\text{-}A,B,C)$評価用データの割合を順序正解率と呼ぶ．

$$\text{順序正解率}(\%) = \frac{\text{正解したデータ数}}{\text{全}(X-ABC)\text{評価用データ数}(1980)}$$

以降の節では，順序正解率を用いて評価を行う．

7. 実験と考察

本論分では，多義分割 CB の作成，および多義性解消関連度計算方式の比較検証を行う．評価には，前節で述べた$(X\text{-}A,B,C)$評価用データを用いている．

多義分割 CB の作成実験について述べる．多義性解消前の CB では，およそ 8 万 7 千語の概念に，平均 30 属性が付与されていた．この CB に対し，4.1.で述べた概念構造情報を基に，概念を一義の概念へと再定義した多義分割 CB では，概念総数がおよそ 8 万 9 千語まで増加し，各概念に対し，平均 28 属性が付与された．

概念：星を例に，多義分割 CB を示す

分割前　星の属性

星，天体，恒星，惑星，点，鋲，犯人，勝ち負け，容疑者，…，歌

分割後　星の「天体」の意味での属性

星宿，星団，星座，恒星，遊星，星雲，エトワール，星占，天体，衛星，公転，星影，太陽系，彗星，星明

分割後　星の「犯人」の意味での属性

罪人，容疑者，下手人，現行犯，お尋ね人，犯，罪業，エトワール，犯罪，自供，情状，捕り物，有罪，侵す

このように，多義分割 CB では，一つの概念：星が持つ複数の意味を定義するために，星という概念を複数定義することによって，意味を分割し扱うことが出来る．すなわち，(犯人，星)の関連度を算出する場合には，星の中でも犯人に最も意味が近い属性を持つ概念と関連度計算をすることによって，正しく演算することが可能となっている．

7.1. 実験

評価実験として以下の 3 項目について順序正解率を求めた．なお，多義分割 CB の評価，多義性解消関連度計算方式の評価に際しては，基準概念 X に対して，高関連と考えられる概念 A を手がかり語として付与し，関連度計算を行っている．すなわち，概念の意味特定のために，概念 A を利用している．

(1) CB と意味関連度計算方式 MR で評価
(2) 多義分割 CB と意味関連度計算方式 MR で評価
(3) CB と多義性解消関連度計算方式 CR で評価

この実験により，多義性を考慮しない状況下での評価，概念ベースを多義分割し多義語を一義の概念に置換して関連度計算を行った評価，多義分割をしない概念ベースを用いて関連度計算方式において動的に属性の対応を決定した評価を行い，各手法の有効性を検証している．

7.2. 結果と考察

7.1.の実験結果を図 2 に示す．図 2 では，CB＋MR の順序正解率に対し，多義分割 CB＋MR の順序正解率のほうがわずかに低い値を示していることが分かる．これは，多義の基準概念 X について，概念 A を基に概念 X の意味を決定し関連度計算を行っているが，正しく意味特定が行われなかった評価セットが存在したこと，また，概念 X と概念 A の関係以外の観点から見て高関連となる概念 C との関連度が意味特定の失敗によって高く算出されたことが原因であると考えられる．

さらに，多義分割 CB の問題点として，意味分類の数が多く，概念を分割しすぎたことによる対応付けの失敗が考えられる．つまり，意味分類の数が増えることにより，概念を定義づける属性が正しく取得できなかったと言える．

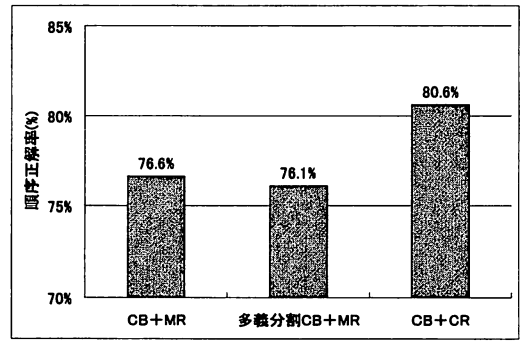

図 2　順序正解率による評価結果

また，CB＋CR の順序正解率は，CB＋MR の順序正解率と比較し 4%向上していることが分かる．これは，多義分割 CB の問題点である意味分類の数が多すぎるために起こる不整合を，意味分類の数を抑制すること

によって，関連度計算により選択的に属性を抽出することによって，正しく概念の意味を推定できたためであると考えられる．以下に正しく意味を推定することが出来た例を示す．

入力多義語：バス　手がかり：ソプラノ

声，混声，四重唱，男声，低音，罵声， 西域，トレーラーバス，テナー，アルト， マイクロバス，ワンマンカー，バリトン， 音域，メゾソプラノ，声部，男声

概念ベースに定義されている概念に付与されている属性は，目視評価によりおよそ7割程度の属性が適切な属性であることが分かっている．そのため，バスの属性から選出された属性のうち，男声に関する属性以外の語も取得されているが，取得された属性のうち，7割(12語/17語)が適切な属性であるため，概念ベースの属性の質とほぼ等価に属性を選出することが出来ていると考えられる．

これらの実験結果により，概念ベースは元来多数の一般的な語に対する関連の強さを定量化するために用いられてきたが，多義性を考慮し，分割を行った概念ベースを用いるよりも，多義性は考慮せずに概念を構築し，関連度計算によって選択的に属性を利用することで，柔軟に語と語の関連の強さを定量化するべきであると言える．

8. おわりに

本論分では，概念ベースの持つ多義性を解消するための手法として，概念分割による概念ベースの再構築手法，ならびに，関連度計算による概念の揺らぎを考慮した属性選択手法を比較検討した．

順序正解率を用いた関連度評価実験により，概念分割によって概念ベースの多義性を解消する手法よりも，関連度計算によって選択的に属性を利用する手法が良好であることを示した．

今後の課題として，本論分で行った評価実験では，多義概念の意味を特定するための語を意図的に付与して関連度計算を行っていたが，入力された語に対して，自動的に手がかりとなる語を付与し，最適な関連度を算出するための手法を検討する必要がある．すなわち，一般的に語と語の関連の強さを定量的に評価するためには，入力された語間に存在する観点を自動的に抽出し，最も関連度の値が高くなる出力を得ることが要求される．また，常識判断メカニズム[5][6]などにおいて使用されている未知語処理など，特定の条件下において多義性を考慮した関連度計算を行う場合においては，各種判断知識ベースの代表的な語を手がかりとして，多義語の意味を特定する手法も必要となるだろう．

また，本論分で比較検討した手法はすべて概念構造情報に基づくものであり，概念ベースに対する新語の学習などにおいては，概念構造情報から意味分類を推定することは難しい．そのため，新語の学習にも対応できる関連度計算方式を検討する必要がある．

謝　辞

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] 広瀬幹規，渡部広一，河岡司，“概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法”，信学技報，TL2001-49，pp.109-116，2002．

[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築法－語間の論理的関係を用いた属性拡張”，自然言語処理，Vol.11，No.3，pp.21-38，2004．

[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.1，2006．

[4] 奥村紀之，渡部広一，河岡　司，“電子化新聞を用いた概念ベースの拡張と属性重み付与方式”，情報処理学会研究報告，2005-NL-166，pp.55-62，2005年3月

[5] 土屋誠司，小島一秀，渡部広一，河岡司，“常識的判断システムにおける未知語処理方式”，人工知能学会誌，Vol.17，No.6，pp.667-675，2002年8月

[6] 吉村枝里子,土屋誠司,渡部広一,河岡司，“連想知識メカニズムを用いた挨拶文の自動拡張方式”，自然言語処理，Vol.13，No.1，pp.117-141，2006年1月

[7] 山西公一郎，小島一秀，渡部広一，河岡司，“国語辞書の意味分類を利用した概念ベースにおける多義概念の分割”，情報処理学会自然言語処理研究会資料，145-6，pp.37-44，2001年9月

[8] 辰己直彦，渡部広一，河岡司，“多義語知識ベースと関連度を用いた多義語の意味理解方式”，信学技報　　，NLC2005-123，pp.55-60，2006年2月