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

2005－ＩＣＳ－139 (28)
2005／3／16

地理に関する常識知識を用いた応答システムの構築

大江 奈緒子† 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: †naoko@indy.doshisha.ac.jp, ‡{watabe, kawaoka}@indy.doshisha.ac.jp

あらまし 本稿では，理科・社会・音楽などの小学・中学などで習う学習知識のモデルの一つとして，地理に関係した自然言語文章の意味を理解し判断を行う「地理常識判断システム」の実現を目指す．そして，地理判断の最終的な目標は日常会話文における意味理解である．常識的な会話とは一つの質問に対して一つの答えを返すことと言えるので，本稿ではまず，地理に関する中学程度の一問一答形式問題の意味理解を行うことを目指す．そのために，地理に関する常識的な会話に必要である知識の構築を行う．地理の問題集からは，問題を言語としてとらえ記号的に処理を行うための知識地理概念ベース，統計資料からは地理的包含関係・位置関係を知識として体系化した地理シソーラスの作成を行う．

キーワード 地理常識判断システム，地理知識ベース，地理概念ベース，地理シソーラス

The construction method of commonsense judgment system for understanding the conversation of geography

Naoko OE† Hirokazu WATABE‡ and Tsukasa KAWAOKA‡

†‡Dept. of Knowledge Engineering and Computer Sciences, Doshisha University

1-3 Miyakodani Tatara Kyotanabeshi Kyoto, 610-0394, JAPAN

E-mail: †naoko@indy.doshisha.ac.jp, ‡{watabe, kawaoka}@indy.doshisha.ac.jp

Abstract This research is aimed at the realization of creating an intelligent robot. In order to make the conversation between intelligent robot and humans possible, it is crucial that the robots have knowledge of topics (i.e. language, mathematics, science, social studies, music, art, etc.) that can be obtained by students in elementary schools. For the first step of this research, the basic conversation system concerning geography was built. The system created in this research can be applied for different topics such as music and art. This system operates under two different geographical knowledge systems. The two knowledge systems created were *the geographic thesaurus* and *the geographic concept database*. With the application of these two knowledge systems a certain level of conversation can be held with the intelligent robot.

Keyword the geographic commonsense Judging System, the geographic knowledge base, the geographic concept database , the geographic thesaurus

1. はじめに

人間は会話中にあいまいな表現や抽象的な表現を受け取った場合にも，連想することにより適切に判断し，会話を続けることができる．これは，言語の意味や概念同士の関係を知識として習得しているためであり，これらは知識であると同時に常識でもある．常識には数量，時間，場所，感覚や感情に関するもの，国語，算数，理科，社会に関するものなどがある．人間の要求を理解できるコンピュータを実現するには，これらの常識をふまえて自然言語文章の意味理解・判断を行うことができる常識判断システムを組み込むことが必要である．本稿では，理科・社会・音楽などの小

学・中学などで習う学習知識のモデルの一つとして，知識を用いて地理に関係した自然言語文章の意味を理解し判断を行う「地理常識判断システム」の実現を目指す．

また，知識について考えると，知識を闇雲に覚えるのは効率が悪い．例えば，「enjoyment」という英単語を知識として持っていなくても「enjoy」の意味を知っていて「-ment」が名詞形を現しているということを知っていれば意味はわかる．「-ment」の知識は「enjoy」だけではなく他の英単語にも適用できる．このように単語を一個一個覚えるよりもある程度の知識とルールを覚えることが常識の獲得には必要である．この知識とルールの構成が英単語の常識と地理に関する常識では違う．地理ならではの知識の構成とルールを構成することで地理に関する常識判断を行い，地理に関する常識的会話を行うことが，研究の基礎である．

そして，地理判断の最終的な目標は日常会話文における意味理解である．常識的な会話とは一つの質問に対して一つの答えを返すことと言えるので，本稿ではまず，地理に関する中学程度の一問一答形式問題の意味理解を行うことを目指す．そのために，地理に関する知識を集めた地理概念ベースや地理シソーラスなどの知識の構築を行い，それを用いて地理に関する意味理解を行う手法を提案し，「地理常識判断システム」を構築する．

2. 地理常識判断システムの対象

まず，「地理」の定義を行う．「地理」とは，地球上の固定された三次元空間にある事物及びそれに関する事象である．これは，小学・中学校で習う「地理」とほぼ同じ範疇にある．

本稿では，「地理」に関する質問文を 1 つ含む問題を地理問題と定義する．また，会話を主眼としたシステムのため，会話に近いと思われる一問一答問題を対象とする．すなわち，扱う地理問題は「小学校 1 年生から 6 年生，中学 1 年生から 3 年生の学習範囲にある地理分野の一問一答問題」である．また，本研究は，研究の第一段階ということもあり，地理分野の中でも特に，行政区分，自然物に関する話題のみを扱うこととする．すなわち，地理問題の中でも産業，農業，工業，歴史，現代社会，時事問題は扱わないこととする．

本研究で対象とする地理問題の例を以下に示す．

- 東南アジア最長の川はなんという川か？
- 中国でもっとも長い川は？
- 北海道から沖縄まで，およそ何 km ですか？
- 日本の都道府県の数はいくつ？

本研究の対象としない地理問題の例を以下に示す．

- 人口増加を抑えることを目的に「１人っ子政策」を実施しているのはどこか？
- 日本は，アラブ首長国連邦やなんという国から，多くの原油を輸入していますか？
- 海岸から２００海里前の水域を何というか？
- 場所によって時刻がちがうことを何と呼ぶ？
- 中国山地などで起こっている人口減少を何という？

また，以降に用いる評価セットは小学・中学問題集から抜き出した問題文 50 問で，同じものを用いる．

3. 地理問題判定

地理に関して常識を持っている，すなわち理解を行っているとは，地理に関する話題（質問文）かどうかの理解も行えて当然である．地理に関する質問文かどうかの判定を行う．地理に関する問題の判定条件として，次の２条件を用いた．

条件①：地名固有名詞が文中に含まれる．

条件②：質問対象語が地理語データベースに存在する．

地名固有名詞とは，「奈良県」のように場所に関して固有の語のことである．地名固有名詞の例を列挙すると「関西地方，アメリカ，アマゾン川，根釧台地，EU，アフリカ，奈良市，北海道，関東平野，日本，世界，北アメリカ」などがある．逆に地理語データベースとは地理に関する一般語が格納されている．これを表 1 に示す．質問対象語[3]とは質問文が答えを求めている語のことで，例えば「世界で一番高い山は？」という文を例に挙げると，質問対象語は「山」となる．質問文から質問対象語を導出するのに，「概念ベースを用いた知的検索における曖昧な質問文の意味理解」[3]を用いる．地理問題判定を評価した結果を表 2 に示す．地理に関する問題は地理に関すると判定すれば正解，音楽と歴史に関する問題に対しては地理に関係ないと判定すれば正解となる．

表 1　地理語データベース

地方,高地,湖,川,市,山脈,半島,大都市,大河,平野

表 2　地理問題判定

	正解	不正解	正解率
地理に関する問題	46	4	92%
音楽に関する問題	49	1	98%
歴史に関する問題	45	5	90%
合計	140	10	93%

4. 知識の作成

地理に関する常識的な会話に必要である知識の構

築を行う．地理の問題集からは，問題を言語としてとらえ記号的に処理を行うための知識地理概念ベース，統計資料からは地理的包含関係・位置関係を知識として体系化した地理シソーラスの作成を行う．

4.1. 地理概念ベース

地理概念ベースとは地理に関しての連想機能を持った概念ベースである．以下に概念ベース[1)]の説明を行う．

ある語 A をその語と関連の強いと考えられる語 a_i と重み w_i の対の集合として定義する．

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

ここで，a_i を 1 次属性と呼ぶ．また便宜上，A を概念表記と呼ぶ．このような属性の定義された語（概念）を大量に集めたものを概念ベースと呼ぶ．ただし，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ものとする．すなわち，属性を表す語もまた概念として定義されている．したがって，1 次属性は必ずある概念表記に一致するので，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．概念ベースは，「概念」が n 次までの属性の連鎖集合により定義されている．地理概念ベースの例（重み省略，1 次属性まで展開）を表 3 に示す．

表 3　地理概念ベース（一部）

概念	属性			
富士山	日本一	高い	山	最高峰
約 60 億人	世界	人口	およそ	
都道府県	日本一	行政区分	47 都道府県	北海道
バチカン市国	世界最小	国	イタリア	面積

また，概念 A と概念 B の関係の深さを定量的にあらわすのが関連度計算[2)]という方法である．それぞれの概念が持っている属性と重みによって関連度計算は行われ，その結果は関連度という数値で表すことができる．関連度は 0 以上 1 以下の連続的な数で表され，関連度が高いものが関連の深い語ということになる．例えば，「富士山」と「日本一」との関連度が 0.426，「沖縄県」と「バチカン市国」の関連度が 0.001 である．

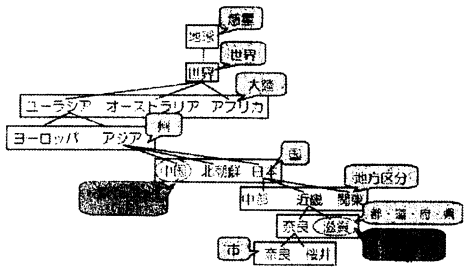
4.2. 地理シソーラス

地理シソーラスは中学用の統計資料を主に用い，手作業で構築した知識である地理シソーラスはツリー構造の地名固有名詞群である．地理シソーラスでは，後述する地理概念ベースでは構築できない地理的包含関係，位置関係を知識として体系化を行ったものである．

地理シソーラスを視覚的に表したものが，図 1 である．ノード“日本”に注目すると，“日本”には，“北海道，近畿，東北，関東など”の地方区分が存在する．さらに“近畿”には“奈良，滋賀，大阪など”の都道府県が存在する．このように地理シソーラスは位置的な包含関係を示すツリー構造になっている．

さらに，階層ごとにその階層に登録されている語の種類を表す行政区分語をつける．図 1 では四角で記されている部分で，「国」，「地方区分」，「都・道・府・県」がこれにあたる．行政区分語の登録によって，固有名詞群の包含関係だけではなく，地理に関する行政区分を含んだ包含関係についても同時に知識として持っていることになる．

図 1 で，一番濃い色で表されているのは山・川・山地・湖などの自然物である．これを「自然語」と呼ぶ．“エベレスト”なら中国とチベットの境にあるので，ノード“中国”とノード“チベット”に付加しており，“琵琶湖”は滋賀県にあるので，ノード“滋賀”に付加している．このように知識体系で格納することで，琵琶湖は滋賀県に存在するだけではなく，滋賀県を包含している近畿地方，さらには日本に存在することを



コンピュータは理解できる．

図 1　地理シソーラス（一部）

図 1 の内部表現形式は表 4 や表 5 のようなデータベースであり，自ノードが「ID」，親ノードが「上位 ID」の上下階層で現されている．地理シソーラスの特徴を以下に列挙する．

- 地理シソーラスは“行政区分語.mdb”，“自然語.mdb”の 2 つのファイルで構成される．
- “行政区分語.mdb”は，「国」，「都道府県」，「州」，「大陸」，「惑星」，「都市」などのテーブルを含んでいる．
- “自然語.mdb”は，「川」，「山」，「山脈」，「海峡」などのテーブルを含んでいる．
- 地名固有名詞に対して一意に ID が与えられている．
- 地名固有名詞はテーブルに記述されている．各テーブルのフィールドは「ID」，「地名固有名詞名」，「上位 ID」，「テーブルが特徴として持っているフィールド」で構成される．
- 「上位 ID」はその地名固有名詞はどこに存在するのかを ID で示したものである．例えば，富士山であれば山梨県と静岡県に存在するので，富士山の上位 ID は山梨県と静岡県の ID が記述されている．

表 4　地理シソーラス（自然語－テーブル　山）

ID	山名	上位 ID	高さ
00000100S	富士山	00000019I	3776
00000101S	白根山	00000019I	3192
00000102S	奥穂高岳	00000020I	3190
00000110S	エベレスト	00000200I	8848

表 5　地理シソーラス（行政区文語－テーブル　国）

ID	国名	上位 ID	首都	面積
00000303I	中国	00000200I	00000603I	9597
00000304I	インド	00000200I	00000604I	3287
00000306I	イラン	00000200I	00000606I	1648

5. 地理常識判断システム

地理概念ベース，地理シソーラスの２つの知識を用いて，地理の一問一答問題の意味理解を行うシステムの作成を行った．このシステムに必要不可欠なのは“富士山は日本で一番高い山”といった人間だけが持っている知識である．従って，本システムでは人手で作成した地理知識ベースという地理に関する知識ベースを用いる（表 6）．地理の一問一答問題を情報文と呼び，情報文を多数格納したものが地理知識ベースである．各情報文には，情報文に含まれる単語が知識（見出し語）として登録されている．

しかし，実際には膨大な形で表記される知識をすべて登録することは困難であり効率が悪い．そこで，代表的な語のみを登録し，地理概念ベースや地理シソーラスにより構築した連想システムを用いて知識の拡張を行い，地理知識ベースに表記の柔軟性を持たせる．すなわち，地理知識ベースに「富士山は日本一の標高を誇る山です」という知識がなくても，地理知識ベースの「富士山は日本で一番高い山です」という知識から連想を行うことができる．地理常識判断システムの流れの概略図を図 2 に示す．

表 6　地理知識ベース（一部）

知識	情報文
富士山①	富士山は日本で一番高い山です.
富士山②	富士山は日本最高峰の山です.
中国①	長江は中国に流れている.
中国②	中国は世界で一番人口が多い.

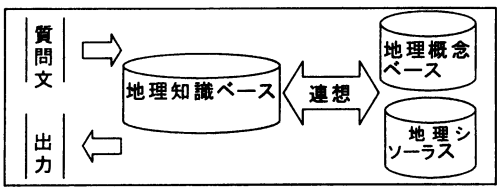


図 2　地理常識判断システム

6. 地理概念ベースを用いた処理

地理概念ベースを用いる方法では，質問文を，自立語の単語列に切り分ける．今，質問文を $\underline{X}$ とすると，次式のように表される．

$\underline{X}=\{(x_1, w_1), (x_2, w_2), \cdots (x_n, w_n)\}$（$x$：自立語，$w$：重み）

地理知識ベース中の情報文も同様に自立語の単語列にする．

$A_1=\{(a_{11}, w_{11}), (a_{12}, w_{12}), \cdots (a_{1n}, w_{1n})\}$

$A_2=\{(a_{21}, w_{21}), (a_{22}, w_{22}), \cdots (a_{2m}, w_{2m})\}$

$\vdots$

質問文 $\underline{X}$ と最も関係の深い情報文を探すために地理概念ベースを用いた文と文の関連度計算を用いる．関連度計算式は以下[4]を用いる．

$$ChainWR(\underline{X}, A)=\sum_{i=1}^{n} MatchWR(x_i, a_u)\times(w_i+w_u)\times(\min(w_i, w_u)/\max(w_i, w_u))\Big/2$$

これにより，最も関連の深い情報文が出力される．

7. 地理概念ベースの有効性

データ 50 問で地理概念ベースを用いたときの処理の評価を行った．正解率は 54%，正解数 27 問となった．不正解の 23 問について不正解理由を見ると，23 問中 10 問が地理知識ベースに解答がない質問文であった．12 問が誤った情報文との関連度が高くなっている．また，1 問はどの情報文との関連度も 0，すなわち，未定語のみで構成された質問文であった．不正解例を以下に示す．

《不正解例》
- 大阪府があるのは，関東地方，近畿地方？
  不正解理由：地理知識ベースに解答となる情報文がない．
- 中国で人口が最大の都市はどこの都市か？
  誤解答：中国の人口は，世界第 1 位である．
  期待する答え：上海は中国最大の人口（870 万人）を持つ都市．
  不正解理由：誤った情報文との関連度が高い．

すなわち，現段階の 571 文の情報では常識知識の不足，追加が必要である．すなわち，地理知識ベースの拡張が必要である．また，関連度計算を用いたことによる有効性を示すために，表記一致（質問文と情報文の一致する自立語の表記一致数が最大の情報文を出力）の結果を表 7 に示す．

表 7　表記一致の評価

	個数
○　一致個数が最高	18
△　一致が最高の群に含まれる（平均出力個数　5 文）	10
×　不正解	22

正解が唯一つ出力されたものが 18 問，正解の他に不正解も出力されたものが 10 問，誤った出力または，

出力がなかったものが 22 問となった．正解のみの出力（○）では地理知識ベースに関連度計算を用いた時（地理概念ベースを用いた処理）の正解数 27 問の方が有効な手法である．また，△の処理においては，平均出力数が 5 文なので，確率まかせに 1 文のみ回答したとしても，20%割合でしか回答が得られない．10 問のテストデータのうち，2 問が正解のみを出力できることとなる．ゆえに，正解が得られるテストデータ数として，18（○）+2（△）=20 問と計算することができ，地理概念ベースを用いた処理の正解数 27 問の方が有効な方法であることが示せた．比較のため，図 3 を示す．

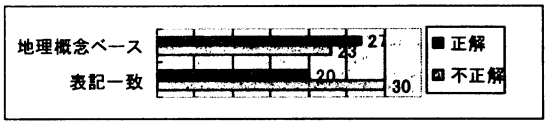


図 3　表記一致と地理概念ベースの評価

地理概念ベースを用いた関連度計算の正解数 27 問の方が有効な方法であることが示せた．

7.1. 地理シソーラスを用いた処理

表の知的検索[5]とは表に対して自然文の質問が与えられたときに，それに対して表から正しい答えを取り出すことである．すなわち，表の知的検索を利用すると，表 4 と「一番標高の高い山は？」という質問から，「エベレスト」という答えが得ることができる．このシステムを地理の回答にも利用する．

表知的検索を用いた処理の流れを，例「日本で一番高い山は？」と合わせて説明を行う．

①質問対象語，条件[3]の取得を行う．

－質問対象語は「山」，条件は「高い」，「一番」である．

②質問対象語もしくは，条件の名前のテーブルが存在するまたは，質問対象語もしくは，条件と高関連度（または，同義語）のテーブルが存在するかどうか．存在しなければ，表の知的検索が利用できないので，地理シソーラスを用いた処理は終了する．

－質問対象語「山」からテーブル「山」を利用する．今回の例では，テーブル「山」を表 5 とする．

③質問文に対して，地名固有名詞の抽出を行う．

－「日本で一番高い山は？」から「日本」を抜き出す．

④　②で選択したテーブルから③で抽出した地名固有名詞のレコードのみテーブルを切り抜く．また，フィールドが ID で記述されているものについては，名称に置き換える．

－表 4 において，「日本」に含まれるレコードを探す．具体的には，上位 ID の上位 ID というように，シソーラスを上位にたどって上位ノードに「日本」が存在すれば，「日本」に存在するレコードということにある．「日本」に存在するレコードのみを切り抜いたテーブルを表 8 に示す．

⑤　④で切り抜いたテーブルと質問文から表の知的検索を行う．

－「日本で一番高い山は？」を表 8 に知的検索を行い，「富士山」という回答を得る．

表 8　表の知的検索に用いるテーブル

ID	山名	上位 ID	高さ
00000100S	富士山	山梨県	3776
00000101S	白根山	山梨県	3192
00000102S	奥穂高岳	長野県	3190

7.1.1. 地理シソーラスの評価

地理シソーラスの評価として，表の知的検索の 7.1 節の④の段階，すなわち，知的検索に表を切り抜いて処理にかける段階までの評価を行った．評価データは質問文 50 問である．正しい表を抜き出せたのは 14 問であった．よって，正解率は 28%である．また，不正解の 36 問には，誤った出力はなくすべて無回答である．次に正解と不正解を示す．

＊正解例

- 中国の人口は約何人？
- 福岡県の県庁所在位置は？
- 日本で最も流域面積の広い川は？

＊不正解例

- 経線・緯線を使った国境の多い大陸は？

不正解理由：地理シソーラスでは回答することができない．

- 岩手県は何地方と同じくらいの広さ？

不正解理由：二つの表データ（県テーブルと地方テーブル）を抜き出す必要があった．現段階ではそのような処理を行っていない．

8. 検討事項

前節でも述べたが，さらに地理知識ベースの拡張が必要になる．その際に関連度計算にかかる実行時間を検討する．結果を図 4 に示す．横軸が情報文の数，すなわち地理知識ベースの規模，縦軸が時間(秒)である．現時点での，地理知識ベースの規模は，情報文数が 571 文で 1.22 秒かかっている．図 4 のグラフを見てもわかるように規模が倍に増えれば，実行時間も倍に増える．したがって，規模が 3 倍になれば，実行時間が約 4 秒かかることになり，人間とコンピュータとの会話においてフラストレーションがたまるだろう．

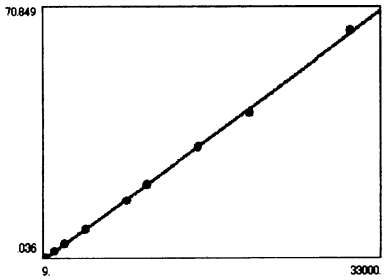


図 4　関連度計算にかかる実行時間

そこで，絞り込み手法を提案する．絞込み手法は単純で，質問文に含まれる知識のみを持つ単語列（情報文）のみを回答の対象とし，候補を絞り込んでから関連度計算を行う．これによって，計算時間がどう短縮されたのか，正解率が変わったのかを表 9 に示す．

表 9　絞込み手法の実行時間と正解率

	知識全体との関連度計算	絞り込んで関連度計算
一問あたりの平均実行時間(秒)	1.25	0.06
正解率(%)	54	54

実行時間が 1/20 ほどになった．これは単純に絞込みによって，回答の候補が 571 文全部と関連度計算とることから 1/20 の約 30 文に絞られたためである．しかし，実行時間の短縮があったのにも関わらず，正解率は変わらなかったため，絞込み手法は有効であると言える．

しかし，今後，地理知識ベースが拡張され，計算時間がかかることが考えられるので，知識の精錬や重要な知識，不要な知識の取捨選択によって，知識ベースが発散しないようにする仕組みを考える必要がある．

9. 地理常識判断システムの評価

入力から出力まで(地理問題判断も含む)の正解率は，地理概念ベースを用いた処理は 52%となった．地理シソーラスを用いた処理は表の知的検索というシステムを処理の内部で用いているのだが，その出力が 100%正しいと仮定すると，正解率 28%となった．

内訳は，50 問の地理に関するテストデータに対して地理問題判定で 4 問の不正解．地理の問題と判断された 46 問に地理概念ベースを用いた処理と地理シソーラスを用いた処理を用いた．地理概念ベースを用いた処理では 26 問の正解が得られた．地理シソーラスを用いた処理では 46 問中，表の知的検索を用いる形にデータを加工できたものが 14 問であった．地理シソーラスを用いた処理からの正解も 14 問になる．

ここで，地理概念ベースを用いた処理と地理シソーラスを用いた処理からの２つの出力から最終出力を得るわけだが，地理シソーラスを用いた処理の出力を優先させて，地理シソーラスを用いた処理で無回答だったものに対して地理概念ベースを用いた処理の出力結果を用いる方法が良い出力結果が得られた．すなわち，地理シソーラスを用いた処理で表の知的検索を用いることのできなかった 32 問について，地理概念ベースを用いた処理の出力を用いると，18 問について正解を得ることができた．よって，地理シソーラスを用いた処理から 14 問，地理概念ベースを用いた処理から 18 問正解を得て，合計 32 問の正解が得られた．地理常識判断システムの最終的な出力に対する正解率は 64%である．

10. おわりに

本稿では，日常会話の中でも特に地理に関係した会話文に着目し，地理に関係する語を集めた知識の構築を行うことを目指してきた．また，知識の評価のための地理常識判断システムを部分的だが構築し，知識の有効性を示すため評価を行った．

そして，日常会話における地理に関する話題においては，より充実した知識が必要であることを示唆した．

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] 眞鍋 康人，小島 一秀，渡部 広一，河岡 司，“概念間の関連度やシソーラスを用いた概念ベースの自動精錬法”，同志社大学，理工学研究報告，Vol. 42，No. 1，pp. 9-20，2001.

[2] 渡部広一，河岡 司，“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol. 8，No. 2，pp. 39-54 ，2001.

[3] 古川成道，渡部広一，河岡司，“概念ベースを用いた知的検索における曖昧な質問文の意味理解”，第 18 回人工知能学会全国大会論文集，2D1-10，2004

[4] 井筒大志，渡部広一，河岡司，“概念ベースを用いた連想機能実現のための関連度計算方式”，情報科学技術フォーラム FIT2002，E-39，pp. 159-160，2002

[5] 権東旭，“連想機能を用いた表理解常識判断システムの構築方式”，同志社大学大学院工学研究科修士論文，2004