

## 自然文知識ベースと連想推論を用いた 質問応答システム

柴山拓也<sup>†</sup> 吉村枝里子<sup>††</sup> 土屋誠司<sup>††</sup> 渡部広一<sup>††</sup>

与えられた質問文に対して応答のために必要な知識を自然文知識ベースから抽出する質問応答システムを構築した。質問応答システムとはコンピュータが人間の与えた教養知識から質問応答のために利用する知識を提示するものである。本研究では理科に関する教養知識を実験的に取上げて研究を行い、完全な自然文で表現された知識を用いた。理科に関する質問文 100 文でシステムの評価実験を行い、出力された知識 3 文のうち 86% が正解となった。

## Question Responding System Using Knowledge Base Built with Natural Language by Associative Inference

Takuya Shibayama<sup>†</sup>, Eriko Yoshimura<sup>††</sup>,  
Seiji Tsuchiya<sup>††</sup> and Hirokazu Watabe<sup>††</sup>

This research aims to realize communication between human and computer for their question responding. Question Responding System which extracts knowledge for question responding from the Knowledge Base built with natural language. The system enables computer to present knowledge for respond to question from traditional knowledge human gives. This study experimentally picks up traditional knowledge of science with natural language. To construct a flexible reasoning mechanism, this research focuses an association that human can infer to knowledge from knowledge base. Evaluating this system by 100 questions of science, it showed 86% accuracy. The Result of this research leads to realize communication between human and computer.

### 1. はじめに

近年、情報化社会がますます発展し、膨大な量の情報を処理できるコンピュータの存在が必要不可欠となっている。コンピュータは人間の生活を便利にし、豊かなものへと変化させてきた。しかし、その一方でコンピュータの複雑化も進み、情報処理技術の知識が乏しい人には扱い難いものとなってきている。そこで、一般家庭や病院、スーパー、福祉施設などといった、人間の生活にとって身近な場所で活躍できる知能ロボットのような、「人間と生活を共にできるコンピュータ」が期待される。そのためには、会話や案内、介護などといった様々な活動の手助けを行える必要があり、多岐に渡る動作が必要不可欠である。コンピュータの動作が日常生活の手助けとなり、誰にでも扱えるようにするために共通して必要な機能は、人間の言葉の理解とその応答、つまりコミュニケーション能力である。日常生活は主に人間同士のコミュニケーションで成り立っており、会話によって他人との意思疎通を図って生活をしている。よって人間と生活を共に出来るコンピュータの実現のためにはコミュニケーション能力が重要であり、人間同士が日頃行っているコミュニケーションをモデル化し、会話を人間とコンピュータのインターフェースにする必要がある。

日常生活にはある人間が質問し、相手の人間がその質問に対して回答するような会話（質問応答）が多く存在する。現在、エキスパートシステムのように人間の質問に対し、特定の分野の知識を用いて回答の行えるシステムが多く研究されている。よく用いられる手法としては、知識を「IF~THEN」のように形式化してデータベースを構築し、質問文を構文解析し表記一致を基本としたデータベース検索により回答を得る方法である[1]。しかし人間の行う質問応答は自然言語のままの知識表現から、単純な形式化では表現しきれない言葉が持つ概念や背景にまつわる知識までを理解する。そして他の関連のある知識を膨らませて推論し、応答に必要な知識を脳内にある膨大な情報の中から適切に抽出することにより、柔軟な質問応答を行うことができる。例えば「塩酸は何性の水溶液ですか？」という質問に対して、まず質問の意味や、「塩酸、性質、水溶液」といったそれぞれの語が持つ概念を理解する。その後「酸性は水溶液の性質である」といった背景にまつわる知識やその他の関連のある知識を膨らませて推論し、「塩酸は酸性である」という応答に必要な知識を脳内より適切に抽出する。最終的にその知識を利用して「酸性」と回答することができる。このように人間は、脳内にある自然文のままの知識から回答となる知識を柔軟に抽出することが可能である。この知識を抽出することができて初めて質問応答が行える。

<sup>†</sup> 同志社大学大学院工学研究科  
Graduate School of Engineering, Doshisha University

<sup>††</sup> 同志社大学理工学部  
Faculty of Science and Engineering, Doshisha University

これまでの手法で用いていた知識文の形式化や表記一致を頼りにした推論では、関連のある知識を膨らますといったような柔軟な質問応答はできない。そこで、自然文による知識の抽出を行える機構を作ることで、人間らしい質問応答を目指す。

人間が様々な表現方法のある自然言語に対応するための機構として「自動車」から「トラック」を想起するような連想が考えられる。このある語から関連のある他の語を想起する連想に着目すれば、語の表記に依存することなく意味を考慮した推論を行うことができると考えられる。そこで本稿では、自然文による質問に対し、応答となる自然文の知識を連想メカニズムを用いて推論する質問応答システムを提案する。そして、完全な自然文を扱う、人間とコンピュータの質問応答を実現する。

## 2. 連想メカニズム

人間は「自動車」から「トラック」や「タクシー」というように、ある語から関連のある他の語を想起することができる。このようなメカニズムを連想メカニズムと呼ぶ。コンピュータは単一の語を独立した情報としてデータベースに保持するため、他の知識との関連性を考慮することができない。よってコンピュータは人間のように連想を行うことができない。そこで以下の技術を用いてコンピュータに連想メカニズムを模倣させる。

連想メカニズムを用いることにより、関連のある知識を膨らませて推論(連想推論)を行うことが可能となり、本研究の目的である人間らしい質問応答を目指す。

### 2.1 概念ベース

概念ベース (CB) とは、国語辞書等から語 (概念) と意味 (属性) のセットを格納し自動構築されたデータベース[2]である。ある単語  $A$  をその単語と関連の強いと考えられる単語  $a_i$  と重み  $w_i$  の対の集合として定義する(式 1)。

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

ここで、 $A$  を概念表記、 $a_i$  を一次属性と呼ぶ。このような (属性, 重み) の対の集合を大量に集めたものを概念ベースと呼ぶ。ただし、任意の一次属性  $a_i$  は、その概念ベース中のある概念表記の集合に含まれているものとする。従って、一次属性は必ずある概念表記に一致するため、さらにその一次属性を抽出することができる。これを  $A$  に対する二次属性と呼ぶ。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。

### 2.2 関連度計算方式

関連度計算方式[3]とは、概念ベースに定義されている概念  $A$  と概念  $B$  の関連の強さを定量的に評価する方法である。各々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 0.0 以上 1.0 以下の連続的な実数で表され、関連度が高いものが関連の深い語ということを示す。例を表 1 に示す。

表 1 関連度計算方式 (概念「自転車」と各概念との関連度)

概念 $A$	概念 $B$	関連度
自転車	自動車	0.322
	飛行機	0.029
	船舶	0.005

## 3. 研究目的

本稿では、与えられた質問文に対して応答のために必要な知識を自然文知識ベースから抽出する質問応答システムについて示す。質問応答システムとはコンピュータが人間の与えた教養知識[4] (教養知識ベース, 教養概念ベース) から質問応答のために利用する知識を提示するものである。教養知識とは「国語」や「社会」といった義務教育で学ぶ知識のことであり、本研究では「理科」を実験的に取り上げて研究を行った。情報源 (人間で例えると知識源となる脳内) としては、知識を文形式で表現した教養知識ベースを用いる。また特徴として、本研究では完全な自然文によって表現された知識 (以降、知識文と呼ぶ) を扱う。これにより今後、知識ベースに格納される知識を増やす際にも、膨大な情報が存在する Web を情報源として利用でき、自動学習[5]によって人手で複雑な処理が不要になるという展望を考えることができる。システムのイメージを図 1 に示す。

図 1 のように、与えられた質問文に対し質問文解析を行い、回答を行うために重要な語を抽出する。それらの語に教養知識と連想メカニズムを用いて、他の関連のある語を膨らませる。質問文解析と連想メカニズムで得られた語を基に推論を行い、知識ベース内から回答のために必要な知識文を抽出し出力する。

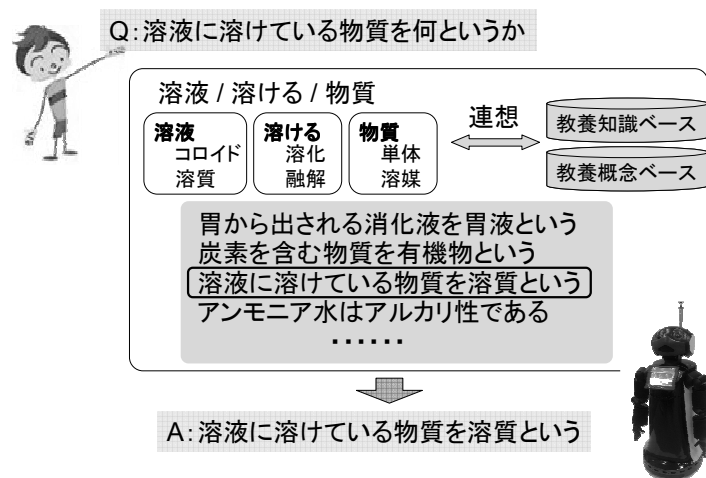


図 1 質問応答システムのイメージ

#### 4. 教養知識

本稿では教養知識を国語や社会といった義務教育（小・中学校）で学ぶ知識と定義する。本研究では「理科」に関する教養知識を実験的に利用しており、例えば「塩酸は酸性の水溶液である」のような知識がこれに当たる。

提案システムにおいて教養知識は二つの知識形態で保持している。質問応答システムで利用する知識文を格納する「教養知識ベース」、教養知識特有の語（例：リトマス紙など）についての連想を可能にする「教養概念ベース」である。

##### 4.1 教養知識ベース

教養知識ベースは自然言語による多種多様な教養に関する質問文の応答のために必要となる知識文を持たせることを目的としたものである。Web や参考書などから知識とすべき文章を抽出し、人手で知識文として格納する。このとき、知識文は今後変わることのない事実のみを対象とする。また教養に関する知識とは関係の無い文章は質問応答で用いないと考えられるため研究の対象外とする。なお、教養知識文は理科に関する 1622 文を実験的に使用し、重複する知識文は存在しない。

##### 4.2 教養概念ベース

教養概念ベースとは、既存の概念ベースに教養に関する固有名詞（ナトリウム、白鳥座など）を追加したものである。国語辞書を基に構成された既存の概念ベースには、教養に関する固有名詞は定義されていないことが多いため、連想や関連度の算出を行うことができない。そこで、既存の概念ベースに教養に関する固有名詞を追加し、教養概念ベースを構築することにより、これらの固有名詞に対しても連想や関連度の算出を行うことを可能にする。

#### 5. 質問応答システム

質問応答システムとは、与えられた質問文の応答のために必要な知識を自然文で表現された知識ベースから抽出し出力するシステムである。システムの全体像を図 2 に示す。

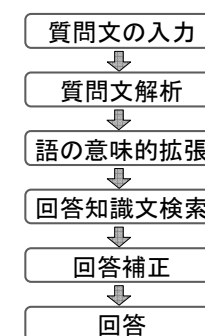


図 2 質問応答システム

まず入力された質問文の解析を行い、質問文の文構造を明確化させる。質問文解析で得られた情報と教養概念ベースによって、質問文中の語と意味的に同じである語（以降、拡張語と呼ぶ）を取得して知識を膨らませた後、知識ベースから回答となる知識文を抽出する。そして、抽出された知識文で質問に対する回答が可能であるかの判定を行う。もし質問に対する回答となり得ない場合は、答えられない部分を答えられるような知識文を抽出して回答補正を行う。最終的に回答である可能性の高い順に順位付けを行い、上位に位置する知識文を回答として出力する。

### 5.1 前提条件

単一の知識文を利用して応答することができる質問文を単数知識問題と呼ぶ。また、二つの知識を必要とし、三段論法を用いて応答することができる質問文を複数知識問題と呼ぶ。この2種類かつ単語単位で回答できる質問文（～は何（色）か？など）をシステムの入力対象とする。

### 5.2 質問文解析

入力される質問文は自然文であるため、そのままでは文中のどの情報（語）が重要かわからない。そこで入力された質問文を構文解析し、質問文にとって重要な情報を取り出す。ここで重要な情報とは質問文の自立語と係り受け情報とする。自立語とは単独で意味を表すことのできる語であり、文章を構成している主要要素と考えられる。また係り受けは文節と文節がどのような関係で繋がっているのかを表しており、文章中でどの文節が重要かを知らず掛かりとなると考えられる。よって、これらの情報を用いて質問文の重要箇所を判別することができる。

まず質問文に形態素解析を行い、自立語を抽出する。次に抽出された自立語に係り受け解析を行い、それぞれの自立語に対して一定の重みを付与する。例えば質問文「水滴が地上の物体に付く現象は何か？」について係り受け解析を行うと図3のような結果を得る。なお、「何」のような疑問詞は自立語であっても質問文に対して重要な情報は得られない不要な語として抽出しない。

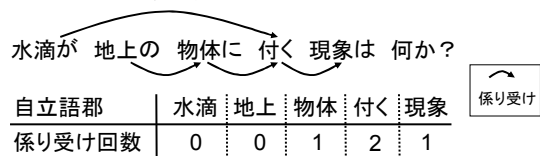


図3 係り受け解析

重みとは文章中の他の語よりも重要度を上げて推論するための処理である。図3のようにそれぞれ自立語の係り受け回数を調べ「係り受け回数×重み係数」をその自立語の重みとする。

### 5.3 語の意味拡張

語の意味拡張とは、教養概念ベースを用いて質問文の語に対し意味的に同じである語（以降、拡張語）を新たに付与する処理である。これにより、語の表記だけでなく、語の意味も考慮した推論が可能になると考えられる。「塩酸」を例に意味拡張の処理を図4に示す。



図4 語の意味拡張

まず教養概念ベースの概念「塩酸」に格納されている属性を重み昇順で獲得する。次に「塩酸」と各属性に対して関連度の算出を行い、閾値以上のものを類義・同義と判断して「塩酸」の拡張語とする。なお、閾値に関しては同義語・類義語リストを用いて実験的に算出した関連度の平均値を用いる。この意味拡張による連想を行うことにより、質問文中の自立語である「塩酸」は拡張語を含め「塩酸、塩化水素、希塩酸」となる。よって、「塩酸」の表記一致のみならず、意味的に関連した推論を行うことができる。本研究ではこの意味拡張を質問文中の自立語全てに行う。例えば質問文「溶液に溶けている物質を何というか」に対し意味拡張を行うと表2のようなになる。

表2 「溶液に溶けている物質は何というか」の意味拡張

自立語	拡張語
溶液	コロイド, 溶質, ...
溶ける	溶化, 融解, ...
物質	単体, 溶媒, ...

### 5.4 知識文検索

本節では、質問文解析で得られた情報と意味拡張で獲得した拡張語を用いて知識ベース内から知識文を抽出する。まず質問文中の自立語と拡張語を用いて知識ベース内から表記が一致する知識文を検索する。次に、検索の際の語の一致度合いを語一致度とし、語一致度と5.2節の重みを加算した値を知識文の点数とする。この値が質問に答えられる知識文である可能性を表す。2種類の語一致度及び知識文の点数について以下に示す。

- 語一致度 A: 質問文の自立語 X 語中、知識文の語に x 語一致。  

$$\text{語一致度} A = \frac{\text{知識文との一致語数}}{\text{質問文の自立語数}} = \frac{x}{X}$$
- 語一致度 B: 知識文の全語 Y 語中、質問文の自立語に y 語一致。

$$\text{語一致度} B = \frac{\text{質問文の自立語との一致語数}}{\text{知識文の全語数}} = \frac{y}{Y}$$

- 点数：回答である可能性を表す値。  
知識文の点数 = 一致度  $A$  × 一致度  $B$  + 重みの総和

表記の一致した全ての知識文に対し点数付けを行い，上位 10 件を回答候補として抽出する。

### 5.5 回答補正

回答補正は回答候補として抽出された知識文のうち，質問文に回答できなかった知識文に対して行う。回答候補である知識文の情報を基に，知識ベースから新たな知識文を抽出し，回答候補と置き換えることを行う。

質問文に回答するための知識文には，質問文中の重要な自立語が多く含まれる必要がある。そこで 5.4 節で示した語一致度  $A$  が低ければ，知識文中に質問文の語が多く含まれず，正解となる可能性が低いと判断する。不正解と判断した回答候補の知識文の情報を基に，新たな知識文を抽出する。新たな知識文を抽出する際の検索には，図 5 のように質問文の語と回答候補である知識文の語との排他的論理和により得られた語とする。

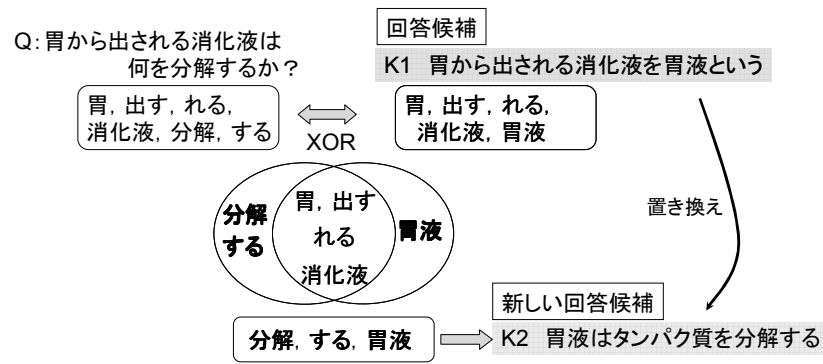


図 5 排他的論理和

排他的論理和で得られた語を基に 5.4 節の手法で知識文を抽出し，その上位 1 位の知識文と不正解と判断された回答候補の知識文を置き換える。最終的に上位 3 件を，与えられた質問に対しての応答として出力する。

## 6. 質問応答文システムの評価

### 6.1 評価方法

評価に使用した質問文は計 100 文（単数知識問題 50 文，複数知識問題 50 文）である。質問文に関しては，あらかじめ教養知識ベース内に質問応答に必要な知識文が存在することを確認し，回答となる語とセットで作成した。本稿では上位 3 件以内に回答として求めている語が含まれている知識文が出力できれば正解とする。質問文形式としては 5.1 節に準じており，質問文と回答となる語の一部を表 3 に示す。

表 3 質問文と回答の例

種類	質問文	回答
単数知識問題	溶液の濃さを何というか	濃度
単数知識問題	大気が物体に及ぼす圧力を何というか	大気圧
複数知識問題	大気の圧力が等しい地点を結んだ線を何線というか	等圧線
複数知識問題	炭素を含む物質が燃えると何ができるか	二酸化炭素

### 6.2 評価結果

評価結果を図 6，単数知識問題 50 文，複数知識問題 50 文での評価結果を図 7 に示す。

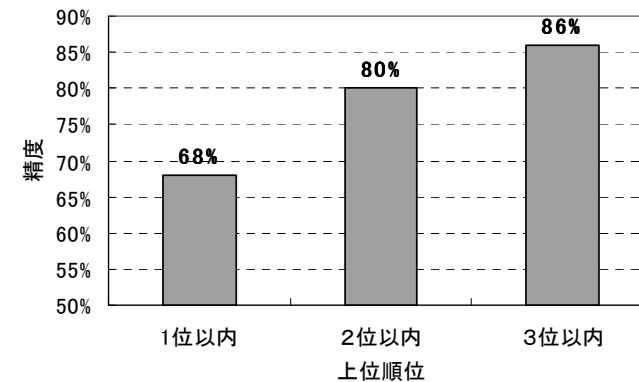


図 6 全体 100 文での評価結果

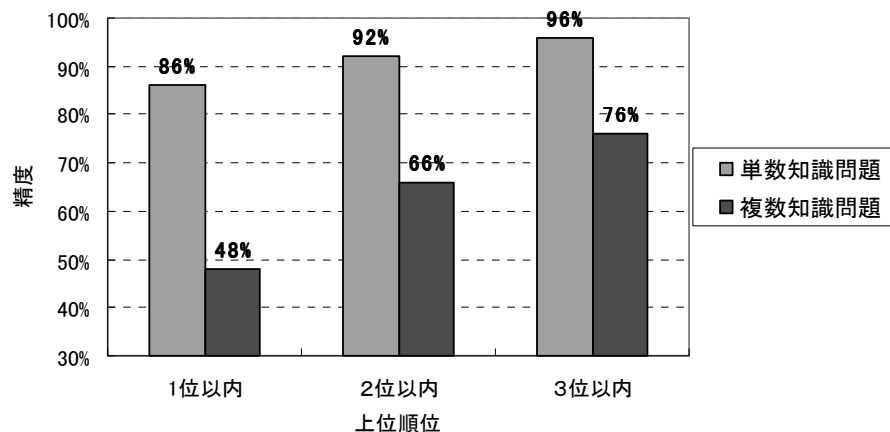


図7 単数・複数知識問題別の評価結果

図6より全体では、上位1位中に正解となる知識文が出力されているのは65%、上位2位中では80%、上位3位中では86%の結果が得られた。

図7より単数知識問題は上位1位で86%、上位2位で92%、上位3位で96%と比較的高い評価結果が得られた。複数知識問題は上位1位で48%、上位2位で66%、上位3位で76%という評価結果が得られた。

## 7. おわりに

本稿では、自然文知識ベースと連想推論を用いた質問応答システムを提案した。質問応答システムとは人手で作成した代表的な教養知識を基に、自然文による質問に回答する知識を推論する。システムの評価実験として質問文100文(単数知識問題50文、複数知識問題50文)を用いた結果、全体としては86%(上位3位以内)を得ることができた。単数知識問題では96%、複数知識問題では74%を得ることができた。全体で86%の精度と、有意なシステムを構築することができ、完全な自然文を用いた人間とコンピュータの教養知識に関する質問応答の実現に近づけることができた。

しかし現在の提案手法のみでは、質問の応答となる知識が上位に抽出されない質問文が複数あることから、さらなる絞込みが必要だと考えられる。今後の課題としては、質問文の解析で、重み付けの手法を改良することが求められる。また複数知識問題の

精度も比較的低いため、新たな推論方法を検討する必要がある。

今後の展望として、さらに幅広い知識の質問応答を実現するためには、人手で作成した教養知識では絶対的な知識量が不足する。そのためコンピュータと人間との教養に関する質問応答は完全には実現できない。しかし、Webから知識の自動学習を用いれば、知識量も自動的に増加させることができる。また、知識というのは一般教養知識のみならず、人間の嗜好などの個人情報なども考えられる。それらを自然言語で与えることにより、教養知識に関する質問のみではなく、日常生活で行われるような様々な会話や質問にも応答でき、完全な人間とコンピュータの会話(質問応答)が実現されると考えられる。

**謝辞** 本研究の一部は、科学研究費補助金(若手研究(B)21700241)の補助を受けて行った。

## 参考文献

- 1) 荒屋真二：“人工知能概論(第2版)”，共立出版，2009
- 2) 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練手法”，信学技報，NLC2001-93，pp.109-116，2002
- 3) 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.39-54，20011
- 4) 中本一志，渡部広一，河岡司：“web情報文からの教養知識の自動学習方式”，信学技報，NLC2007-93，pp.33-37，2007
- 5) 日下裕介，渡部広一，河岡司：“Web上の大容量知識を用いた教養知識の自動学習方式”，情報処理学会研究報告，2009-NL-190(5)，Vol.2009，No.36，pp.29-34，2009