E-001

電子化国語辞書の知識に基づく自然言語質問文応答方式

The Reply Method for Natural Language Question based on Knowledge of an Electronic National Language Dictionary

日下　裕介†　　渡部 広一†　　河岡 司†
Yusuke Kusaka　Hirokazu Watabe　Tsukasa Kawaoka

1．はじめに

近年，コンピュータは人間の道具として非常に便利な存在であるが，今後，人間のパートナーとしての役割がこれまで以上に期待されると考えられている．そこで，人間とコンピュータの円滑なコミュニケーションを取れる手法が必要とされている．人間同士のコミュニケーションにおいて重要な役割を果たすものの 1 つとして会話が挙げられる．人間は会話の中で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，言語の意味や語（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．人間が発する会話の意図を理解できるコンピュータを実現するには，この常識をふまえて自然言語文章の意味理解・判断を行うための知識とメカニズムを構築することが必要である．

本稿では電子化辞書／辞典の大量知識を用い，自然言語による言葉の意味知識に関する質問文に解答する．そのため連想メカニズムである概念ベース[1]や関連度計算方式[2]，また情報文を知識として格納した知識ベースを使用し，知識文を自動的に抽出する手法を提案する．また評価実験により提案システムの有効性を示す．従来の質問文応答システム[3]は，絞り込み型質問応答システムを提案している．本論文では絞り込み型質問応答システムの問題点を明確にするとともに，返答システムの精度向上を目指す．それにより人間がコンピュータに「ワインの材料は何ですか？」と質問した場合，「ワインは葡萄の果汁で作られたお酒です」というようにコンピュータが自然な返答文を返すことが可能となり，今後の会話システムの発展に貢献できると考えている．

2. 連想メカニズム

2.1. 概念ベース

概念ベースは，電子化された複数の辞書から抽出した概念表記や属性によって機械的に構築され，約 9 万語の概念を蓄えた知識ベースである．概念は，ある語 A をその語と関連の強いと考えられる語（属性）a_i と重み $w_i(>0)$の対の集合として定義する．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (2.1)$$

ここで，属性 a_i を概念 A の 1 次属性と呼ぶ．また，属性 a_i も概念ベースに登録されている 1 つの概念である．従って，a_i からも同様に属性を導くことができる．a_i の属性 a_{ij} を概念 A の 2 次属性と呼ぶ．

†同志社大学大学院工学研究科
Graduate School of Engineering Doshisha University

2.2. 関連度計算方式

関連度計算方式は，概念ベースに定義された語と語の関連の強さを，同義性，類似性のみに関わらず定量化する手法である．以下，概念間の一致度，並びに一致度に基づき関連度を求める関連度計算方式について述べる．

2.2.1　一致度

概念 A，B の属性を a_i，b_j，対応する重みを u_i，v_j とし，それぞれ属性が L 個，M 個あるとする($L\leqq M$)．また，各概念の属性の重みを，その総和が 1.0 となるよう正規化している．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_L,w_{Ln})\} \quad (2.2)$$

$$B=\{(b_1,v_1),(b_2,v_2),\cdots,(b_M,v_M)\} \quad (2.3)$$

このとき，概念 A と概念 B の属性一致度 $MatchWR(A,B)$ を以下のように定義する．ただし，$a_i=b_j$ は属性同士が一致した場合を示している．すなわち，一致した属性の重みのうち，小さい方の重みの和が一致度となる．また，一致度は 0.0~1.0 の値をとる．

$$MatchWR(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha\ (\alpha\leq\beta)\\ \beta\ (\alpha>\beta)\end{cases} \quad (2.4)$$

2.2.2　関連度

概念関連度 MR は，対象となる二つの概念において，一次属性の組み合わせについて一致度を求め，これを基に概念を構成する属性集合全体としての一致量を計算することで算出される．

具体的には，見出し語として一致する属性同士($a_i=b_j$)について，まず優先的に対応を決定する．他の属性については，全ての一次属性の組み合わせにおいて属性一致度を算出し，属性一致度の和が最大となるように組み合わせを決定する．一致度を考慮することにより，属性同士の見出し語としての一致だけではなく，一致度合いの近い属性を有効に対応づけることが可能となる．また，概念 A，B 間の見出し語として一致する属性($a_i=b_j$)については，以下の処理により別扱いとする．$a_i=b_j$ なる属性があった場合，それらの属性の重みを参照し，$u_i>v_j$ となる場合は，a_i の重み u_i を u_i-v_j とし，属性 b_j を概念 B から除外する．逆の場合は，同様に b_j の重み v_j を v_j-u_i とし，属性 b_j を概念 B から除外する．見出し語として一致する属性が T 組あった場合，概念 A，B はそれぞれ A'，B' として以下のように定義し直され，これらの属性間には見出し語として一致する属性は存在しなくなる．

$$A'=\{(a'_1,u'_1),(a'_2,u'_2),\cdots,(a'_{L-T},u'_{L-T})\} \quad (2.5)$$

$$B'=\{(b'_1,v'_1),(b'_2,v'_2),\cdots,(b'_{M-T},v'_{M-T})\} \quad (2.6)$$

見出し語として一致した属性の関連度を *MR_com(A,B)* とし，以下の式で定義する．

$$MR\_com(A,B)=\sum_{a_i=b_j}\min(u_i,v_j)$$
$$\min(\alpha,\beta)=\begin{cases}\alpha\ (\alpha\le\beta)\\ \beta\ (\alpha>\beta)\end{cases} \quad (2.7)$$

次に，見出し語として一致する属性を除外した *A'*，*B'* の関連度を *MR_def(A',B')* とする．*MR_def(A',B')* を算出するために，属性数の少ない方の概念 *A'* の並びを固定し，属性間の属性一致度の和が最大になるように概念 *B'* の属性を並べ替える．このとき，対応にあふれた属性は無視する．概念 *A'* の属性 a'_i と概念 *B'* の属性 b'_x が対応したとすると，概念 *B'* は以下のように並び換えられる．

$$B'=\{(b'_x,v'_x),(b'_{x+1},v'_{x+1}),\cdots,(b'_{x+L-T},v'_{x+L-T})\} \quad (2.8)$$

そして，見出し語として一致する属性を除去した属性間の関連度 *MR_def(A',B')* を以下の式で定義する．

$$MR\_def(A',B')=\sum_{s=1}^{x+L-T}Match(a'_s,b'_s)\times\frac{\min(u'_s,v'_s)}{\max(u'_s,v'_s)}\times\frac{u'_s+v'_s}{2}$$
$$\min(\alpha,\beta)=\begin{cases}a(\alpha\le\beta)\\ \beta(\alpha>\beta)\end{cases},\ \max(\alpha,\beta)=\begin{cases}\alpha(\alpha\ge\beta)\\ \beta(\alpha<\beta)\end{cases} \quad (2.9)$$

このように，見出し語として一致する属性間の関連度 *MR_com(A,B)* と，それら以外の属性間の概念関連度 *MR_def(A',B')* をそれぞれ算出し，合計を概念 *A*，*B* の関連度 *MR(A,B)* とする．

$$MR(A,B)=MR\_com(A,B)+MR\_def(A',B') \quad (2.10)$$

関連度も，一致度と同様 0.0~1.0 の値をとる．

3. 知識ベース構築方法

電子化辞書からの知識獲得方法の説明を行う．辞書説明文の一文ごとに独立した一つの知識として取得する．そして取得した説明文を句点“。”区切りで分割し，一つの知識として獲得する．その際，辞書の見出し語となっている語を，知識の索引語として与える．図 1 は辞書中の“鉄道”に対する説明文であり，表 2 は図 1 より獲得した国語知識ベースの例であり，知識文が格納されている．これを知識文パートとし出力する応答文として使用する

> レールを敷設した線路上で動力を用いて列車を運転する施設。一六世紀イギリスの鉱山で鉄板を敷いた上に馬車を走らせた。

図 1　“鉄道”の説明文

表 2　国語知識ベース（知識文パート）

索引語	知識文
鉄道	レールを敷設した線路上で動力を用いて列車を運転する施設
鉄道	一六世紀イギリスの鉱山で鉄板を敷いた上に馬車を走らせた

表 3 は表 2 で示した知識文を茶筌[4]により形態素解析し，自立語に分けた状態で格納したものであり，これを INDEX パートとする．この INDEX パートを同時に取得し，関連度計算や検索に使用する．

表 3　国語知識ベース（INDEX パート）

索引語	知識文自立語
鉄道	レール　敷設　線路　上　動力　用いる　列車　運転　施設
鉄道	世紀　イギリス　鉱山　鉄板　敷く　上　馬車　走る

入力された質問文と知識文をそれぞれ概念としてとらえ，それらの自立語を一次属性とし 2.2 節で示した関連度を計算する．そして質問文との関連度が高い知識を回答する．

4. 絞り込み型質問応答システム

4.1. 索引語検索と関連度計算

絞込み方質問応答システムでは索引語検索と関連度計算によって知識検索を行う．知識ベース中の文章は約 240000 文あり，全知識文との関連度計算は膨大な時間がかかり現実的でない．そこで索引語検索により，ある程度質問文と関連があると思われる知識に絞込み，それらと質問文との関連度計算を行う．質問文応答の流れを質問文“鉄道が運搬するものは”という例を用いて説明する．その流れを図 2 に示す．

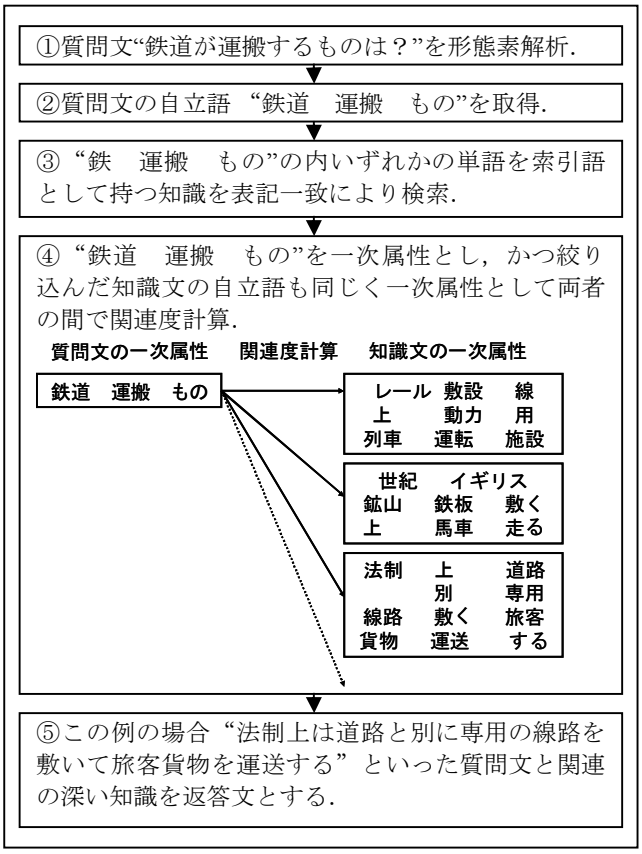

図 2　質問文応答の流れ

4.2. 絞り込み型質問応答システムの問題点

絞り込み型質問応答システムでは関連度計算の回数を減らすため知識の絞込みに索引語検索を使用している．しかし索引語検索では，絞り込み段階において正解となる知識が脱落する可能性が考えられる．例えば，“時刻

を示す機械は？”というような質問文の場合である．その返答文として期待する知識は索引語が“時計”であり知識文が“時刻を示したり時間を計ったりする機械”というものである．例として索引語が“時計”である知識文を表 4 に示す．

表 4　索引語が“時計”である知識文

索引語	知識文
時計	時刻を示したり計ったりする機械
時計	時刻を示しまたは時間を測定する器械
時計	振子の等時性を利用して，指針を等時的に進ませる装置から成る

このような場合知識文の索引語である“時計”は質問文の答えであるために，質問文の自立語を索引語から検索する絞り込み型質問応答システムでは知識文を検索することができない．このように，質問文の答えが知識文の索引語となっている場合は絞込み段階において正解となる知識が脱落する．これが絞り込み型質問応答システムの問題点である．

5. 全体検索型質問応答システム

5.1. 全体検索

索引語検索の問題点の解決方法として全体検索を提案する．全体検索とは索引語からのみの検索ではなく，知識文の自立語からも検索することにより答えとなる知識文を絞込みで残す可能性を高めるという手法である．例えば“時刻を示す機械は？”という質問文の答えとして索引語“時計”，知識文“時刻を示したり時間を計ったりする機械”というものである．その場合，質問文を形態素解析し，得られる自立語は“時刻　示す　機械”であり，知識文の自立語は“時刻　示す　時間　計る　する　機械”である．この状態で索引語検索では，質問文の自立語が索引語に無いため，知識を脱落させてしまうが，全体検索では索引語のみではなく質問文の自立語と知識文の自立語からも検索し，一つでも一致すれば，絞込みで残すというものである．この方法であれば，質問文の答えが索引語であっても答えとなる知識文を脱落させることはない．

5.2. 返答文重要度

全体検索では多くの関連度計算をするため計算時間が増える．そこで，返答文重要度を使用し，関連度計算の回数を減らす．

返答文重要度とは質問文とより情報が一致する知識を重要視し、その度合いを示す値である．

質問文として“鉄道の運搬するものは？”というものがあるとする．これを形態素解析し得る自立語は“鉄道　運搬　もの”である．“もの”は多義語であるので検索はしない．よって“鉄道　運搬”を使用する．また，表記揺れにも対応させるため，概念ベースを参照しこれらの語の属性を展開し，これらの語との間で関連度を計算する．そして関連度が高く算出された属性を拡張する．質問文の自立語“鉄道　運搬”を拡張した語は“線路　列車　レール　運送　貨物”となる．そしてそれらの語を索引語と知識文の自立語より表記一致で検索し，知識文に返答文重要度をつける．開始時はすべての知識文の返答文重要度は 0 であるが，質問文の自立語が 1 つある場合には+2，また拡張した自立語 1 つがある場合は+1 を行うことにする．表 5 に質問文“鉄道の運搬するものは？”に対する返答文重要度を追加した国語知識ベースの例を示す．

表 5　返答文重要度を追加した国語知識ベース

索引語	知識文自立語	返答文重要度
鉄道	レール 敷設 線路 上 動力 用いる 列車 運転 施設	5
鉄道	世紀 イギリス 鉱山 鉄板 敷く 上 馬車 走る	0

返答文重要度の値が高い知識文ほど質問文の答えである可能性が高いと考え，値の高いものから順に関連度計算を行う．返答文重要度を基に関連度計算の回数を指定することにより，精度を保持しつつ計算時間を短縮することができる．関連度計算の回数は実験により閾値を定める．表 6 に“鍔の無い短刀は？”という質問文に対する返答文重要度とその知識文数の例を示す．

表 6　返答文重要度と知識文数の例

返答文重要度	知識文数
1	971
2	21510
3	131
4	2059
5	26
6	346
7	5
8	62
9	0

表 6 からわかるように，返答文重要度が 2 の場合知識文数が 21510 文と莫大な数になる．そこで質問文と知識文との関連度計算の回数を制限するための閾値を用いる．返答文重要度の大きい方から順に知識文数を調べ，知識文数が閾値以下なら関連度計算を行い，次の返答文重要度を調べる．もし知識文数が閾値以上ならそこで関連度計算を終了する．例えば表 6 のような場合，閾値を 2000 とした場合返答文重要度が 5～9 までの知識文を関連度計算し，閾値が 3000 の場合は 3～9 までの知識文を関連度計算するということになる．

6. 評価

6.1. 評価方法

評価を行なうために研究室 40 人の学生からアンケートによって収集した，言葉の意味に関する常識的な合計 102 個の質問文を用いて評価を行なう．評価に使用した質問文の一例を表 7 に示す．

表 7　質問文の一例

ワインの材料は？
列車が運搬するものは？
相手を威圧する態度をなんと言いますか？
鍔の無い短刀は？
稲を食べる害虫は何ですか？
僧が守らなければならない規律は？

これら質問文を入力文とし，質問文と関連が高く算出された上位 20 件の知識分を回答として適切か目視で評価する．上位 20 件に返答文として適切な知識文があれば正解，なければ不正解とする．

6.2. 閾値の評価結果

閾値の検証をするため，閾値を 100～5000 まで 100 ずつ変化させ精度と平均計算時間を求めた．その評価結果を図 3 に示す．

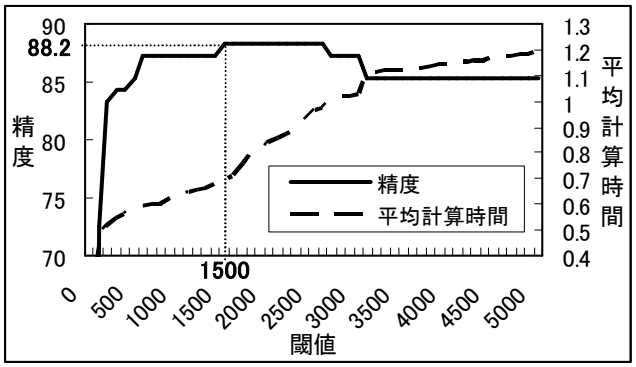

図 3　閾値の検証

閾値を 1500～2600 に設定した場合の精度が 88.2%と一番高いものとなった．6.3 の評価結果では 1500～2600 の中で一番平均計算時間が短い 1500 を採用した．図 3 では閾値を 2700 に設定すると精度が 87.3%，3100 に設定すると 85.3%というように精度が下がった．これは単純に知識文関連度計算する知識文の量を増やすと，答えとなる知識文以外に関連度が高いとされる知識文が現れる可能性がが増えるためである．そのためただ閾値を大きくし関連度計算する知識文の量を増やすのではなく，精度と計算時間から適切な閾値を設定する必要がある．

6.3. 各手法の評価結果

索引語検索，全体検索，返答文重要度（閾値 1500）の評価結果を図 4 に示す．

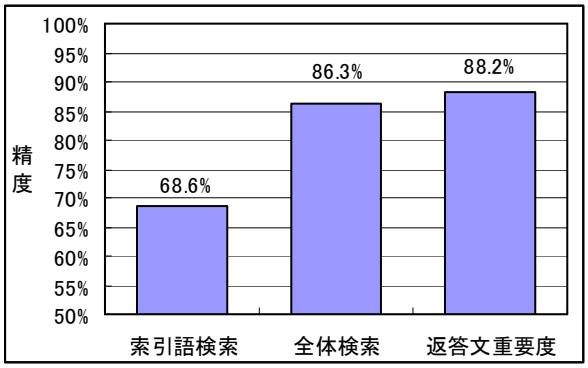

図 4　各手法の評価結果

返答文重要度を使用した場合，全体検索よりも精度が高くなった．膨大な量の知識文と質問文を関連度計算した場合，返答文として適切でない知識文でも返答文となる知識文より高い関連度が出る可能性が増える．そこで返答文重要度で知識文を効率的に絞り込むことによって関連度計算する知識文は洗練されたものとなり，より答えとなる知識文を返答することができた．

次に，各手法で質問文 1 文に返答するまでの平均計算時間を表 8 に示す．

表 8　平均計算時間

	平均計算時間
索引語検索	0.46 秒
全体検索	3.69 秒
返答文重要度	0.69 秒

全体検索は索引語検索より精度が高いが，計算時間が遅くなる．しかし，返答文重要度を使用することによって索引語検索に近い計算時間で，且つ全体検索よりも高い精度となったため，返答文重要度は有効的な手法であるといえる．

7. おわりに

本稿では電子化辞書／辞典から構築した国語知識ベースを用い，自然言語による言葉の意味知識に関する質問文に回答するために関連度を用いて複数の知識文を自動的に抽出する手法を提案した．また，その手法は全体検索や返答文重要度を使用し，絞り込み型質問応答システムより高い精度を実現することができる．このようにコンピュータが人間との質問応答を可能にすることによって，人間と円滑なコミュニケーションを可能にし，人に優しく誰もが利用できる知的コンピュータが実現可能となる．

8. 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．

参考文献

[1] 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－属性信頼度の考え方に基づく属性重みの決定，自然言語処理,Vol.9, No.5 pp.93-110,2002

[2] 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.53-74，2006

[3] 岡田信哉，渡部広一，河岡　司：電子化辞書／辞典を用いた言葉に関する意味知識の自動抽出方法，信学技報，NLC2005-124，pp.61-66，2006

[4] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座，http://chasen.naist.jp/hiki/ChaSen/，2003