

意味処理に基づく文からの回答生成方式

A Method of Generating Answer by Processing Meaning from Sentence

東 宏樹† Hiroki Azuma 渡部 広一† Hirokazu Watabe 河岡 司† Tsukasa Kawaoka

1. はじめに

近年のコンピュータは高性能化が進み、人間の日常生活において必要不可欠なものとなったが、一方で複雑化も進み、コンピュータに関する知識の無い人間にとっては使い難いものとなっている。そのために、人間のコミュニケーションの仕組みをモデル化し、コンピュータとのインターフェイスにもこれを取り入れることが必要となる。人間とのコミュニケーションを可能とするコンピュータの実現のためには、人間と同じ判断能力、そして会話能力をコンピュータに持たせることが必要不可欠だと考えられる。そのために、本稿では、質問文意味理解システム^[1]と意味理解システム^[2]、また、現在構築中である知識フレームなどを用いることにより、質問の意味を理解し、事前に与えられている文からの確かな答えを導き出す質問文意味解釈システムの構築を行う。

2. 関連技術

2.1 概念ベース^[3]

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。

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

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 (87242 語) 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^[4]

関連度とは、定義された知識の関連しか特定できない意味ネットワークのようなものとは違い、任意の概念と概念の関連の強さを定量的に評価するものである。具体的には概念連鎖により概念を 2 次属性まで展開したところで、最も対応の良い 1 次属性同士を対応付け、それらの一致する属性個数を評価することにより、関連の強さを定量化するものである。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 に例を示す。

表 1 関連度計算の例

概念 A	概念 B	概念 A と概念 B の関連度
自動車	車	0.920
自動車	馬	0.031

2.3 質問文意味理解システム

質問文が何を求め、そこにどのような条件が含まれているかを知ることができれば、質問文の内容を理解できたとと言える。質問文意味理解システムに質問文を入力すると、質問対象語 (質問文が求めている対象) とその尺度 (質問対象語「価格」の場合は尺度「通貨の価値」など)、その条件 (質問対象語にかかっている条件)、疑問詞を抽出することができる。例えば、「昨日買った林檎は何個あるか?」という質問文を入力すると、質問対象語は「個数」で、尺度は「数量」、条件は「昨日買った」、疑問詞は「WHAT」となる。

2.4 意味理解システム

意味理解システムとは、単文入力に対して [主体], [何], [時間], [場所], [方法], [理由], [誰に], [用言] という 8 個の成分で構成される意味フレーム (図 1) に分けることができるシステムである。

「昨日 本棚に (1冊 260gの本) が 20冊あった」

主体	何	時間	場所	方法	理由	誰に	用言
1冊 260gの本		昨日	本棚				20冊ある

図 1 意味フレームの例

このように、入力文を意味フレームという一定の形式に変換することで、文型を限定しない形式での入力に対応することができるようになる。

2.5 知識フレーム

知識フレームとは知識ベースや外部から得た情報文の知識的な意味を解析し、また構築した学習知識を参照しフレームに格納するものである。表 2 に知識フレームの格納例を示す。

例) 徳川家康は江戸に幕府を開いた

表 2 知識フレームの例

人物	用語	物	時間	場所	用言
徳川家康	幕府			江戸	開いた

3. 質問文意味解釈システム

3.1 質問文の意味解釈

質問文の意味解釈とは与えられた質問文に対して、情報文から回答を導き出すということである。例えば、「大阪府は大阪府の県庁所在地である。」という情報文がすでに与えられているとする。そこで、「大阪府の県庁所在地はどこですか?」という質問文に対し、情報文から答えを探し、「大阪府です。」という返答をすることを質問文の意味解釈と定義し、この処理を可能とするシステムが質問文意味解釈システムである。ただし、本システムにおける情報文とは単文の知識文のことであり、質問文は情報文の中に必ず答えがあるという前提の上で作成した。なお、知識文とは、「徳川家康は江戸に幕府を開いた」のようなある分

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

野(例文の場合だと歴史)に関する知識に関係する文のことである。また、本システムは、2章で述べた意味理解システム、質問文意味理解システム、知識フレームを利用して、知識フレームは現在構築中のシステムであるため、知識フレームには情報文の情報が正しく格納されたという前提の上で、本システムを構築した。本システムの流れを図2に示す。

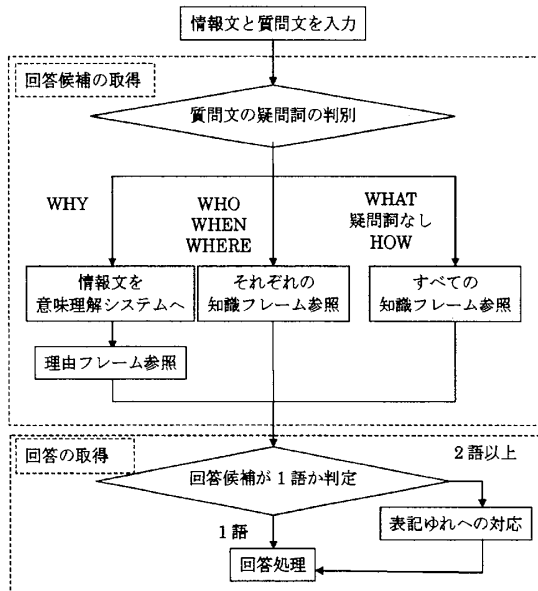


図2 質問文意味解釈システムの流れ

意味理解システムと知識フレームを用いることにより、質問文意味理解システムによって理解できた質問文に対して、その質問に対するより正確な回答が可能となる。意味理解システムでは情報文を文法的に理解し、知識フレームでは各自立語の意味を理解することができる。よって、幅広い形式の質問文や情報文からの回答が可能となり、そこで得た情報から回答を絞り込むことによって、あらかじめ持っている膨大な知識の文の情報から、最も正解に近い回答を出すことができる。

このシステムの全体像は、コンピュータがさまざまな情報文を持っており、そこで質問文を入力すると、そのコンピュータが持っている情報文の中から、その質問文に最も関連の強い情報文を取り出す。その取り出せた情報文から意味理解システムと知識フレームを用いて回答する。このような流れになっている。そこで、質問文に最も関連の強い情報文を取り出す部分と知識フレームは現在構築中のため、正しく情報文を取り出し、知識フレームに格納されたと仮定し、1つの情報文と1つの質問文が与えられたときに回答できるシステムを構築した。

3.2 情報文と質問文の入力

情報文の処理については、知識フレーム、または意味理解システムを用いて、情報文の整理を行う。

質問文の処理については、質問文意味理解システムを用いて、質問文から「質問対象語」と「その条件」、「疑問詞」を取得する。

3.3 回答候補の取得

3.2節で取得した語を利用して、回答候補を導き出す。以下に回答候補を導く方法を質問文の疑問詞別に説明する。

3.3.1 疑問詞別の処理(1)

疑問詞が「WHO」、「WHERE」、「WHEN」の場合の処理を説明する。この場合、それぞれ知識フレームの「人物[成分]」、「時間[成分]」、「場所[成分]」を参照する。その際、参照した[成分]内に語が1語しかなければ、その語を回答として出力する。

[成分]内に語が2語以上ある場合、その中から回答を絞り込む必要がある。参照した[成分]内の語と、質問対象語とその条件をそれぞれ比較し、一致した語は回答にはふさわしくないと判断でき、回答候補を絞り込むことができる。例を図3に示す。

情報文: 「オーストラリアは日本の真南に位置する。」
 質問文: 「オーストラリアはどこかの真南に位置しますか?」

人物	用語	物	時間	場所	用言
	真南			オーストラリア	位置する
				日本	

質問対象語: 場所 条件: オーストラリア, 真南, 位置
 疑問詞: WHERE
 回答候補: 日本

図3 回答候補の取得例

3.3.2 疑問詞別の処理(2)

疑問詞が「WHAT」、「疑問詞なし」の場合の処理を説明する。この場合、知識フレームのどの[成分]を参照すればよいかわからない。また、意味フレームの何[成分]も回答にはふさわしくない語が存在することが多かったため参照しない。そこで、知識フレームの用言[成分]以外から絞り込む処理をする。ここで用言[成分]を対象としないのは、「WHAT」もしくは「疑問詞なし」で問われた場合、用言を含む回答はほとんど無いと考えられるからである。ここでも、「質問対象語」と「その条件」を利用して回答候補の絞り込みを行う。例を図4に示す。

情報文: 「近畿地方の中央低地には日本最大の面積の琵琶湖がある。」
 質問文: 「近畿地方の水瓶と呼ばれる、日本一の面積を誇る湖水を何と言いますか?」

人物	用語	物	時間	場所	用言
	面積			近畿地方	ある
				中央低地	
				琵琶湖	

質問対象語: 湖水
 条件: 近畿地方, 水瓶, 呼ぶ, 日本一, 面積, 誇る
 疑問詞: WHAT
 回答候補: 中央低地の琵琶湖

図4 回答候補の取得例

3.3.3 疑問詞別の処理(3)

疑問詞が「HOW」の場合の処理を説明する。この場合、知識フレームのどの[成分]を参照すればよいかわからない。また、意味フレームの方法[成分]も回答にふさわしくない語が存在することが多く、参照しない。そこで知識フレーム内の語から絞り込む処理をする。疑問詞が「HOW」の

場合は、用言を含む回答が問われることも多く、用言[成分]も絞り込む対象とする。ここでも、「質問対象語」と「その条件」を利用して回答候補の絞り込みを行う。また、用言[成分]内の語が回答に残った場合、用言[成分]には動詞の基本形で格納されるので、回答する際には情報文の状態に直す処理を行う。例を図5に示す。

情報文：「当時、キリスト教とその信者はキリシタンとよばれていた。」
 質問文：「当時、キリスト教やその信者はどうよばれていたか？」

人物	用語	物	時間	場所	用言
	キリスト教		当時		よばれる
	信者				
	キリシタン				

質問対象語：方法
 条件：当時、キリスト教、信者、よばれる
 疑問詞：HOW
 回答候補：キリシタン

図5 回答候補の取得例

3.3.4 疑問詞別の処理(4)

疑問詞が「WHY」の場合の処理を説明する。この場合は、意味理解システムを利用して回答する。意味理解システムの理由[成分]を参照し、その情報を回答する。

3.3.5 表記ゆれへの対応

(1) 関連度計算の利用

知識フレーム内の語と、質問対象語とその条件とを比較し、表記一致しなかったとき、語が複数残る。このような場合、残っている知識フレーム内の語と、質問対象語とその条件の関連度を計算し、閾値 0.36 (高関連度) 以上であれば、その語は同じ意味であると判断し、求める回答ではないと考える。

この処理を行うと次のような問題が起こる。「どんなものですか?」といったような、疑問詞「HOW」が名詞に係る問い方が存在するが、この場合、疑問詞に係っている名詞を比較の対象とさせると、回答候補の中の、回答に必要な語まで回答候補から削除されてしまうことがある。そのため、知識フレーム内の語との比較には、疑問詞に係っている名詞を省く処理を行っている。

(2) 関連度計算が利用できない場合

人名に関しては、同一人物を表している語同士であっても高関連とならない、または計算が行えない場合がほとんどである。そこで、人物[成分]に語がある場合、質問対象語とその条件を形態素解析で区切り、人名と判断された語と人物[成分]内の語が一致した場合、同一人物と判断させる処理を行っている。

3.3.6 回答処理

前節までの処理の結果、得られた回答候補が1つであった場合は、それを回答として出力する。しかし、回答候補が複数残っている場合、回答候補をつなぐ処理を行う。残っている回答候補と情報文を比較させることにより、回答候補をつないで出力する。例えば、情報文「バルカン半島は当時、ヨーロッパの火薬庫とよばれていた。」に対して「バルカン半島は当時どのようによばれていたか?」という質問文を与えれば、「ヨーロッパ、火薬庫」と回答候補

に存在する。その語をつなぐために情報文と比較し、「ヨーロッパの火薬庫」と回答候補をつないで出力する。

4. 評価

質問文意味解釈システムを構築し、質問文に対して的確な回答ができるか評価を行った。情報文は、音楽、美術、地理、歴史に関する知識文を計 96 文用意し、その一文ずつに対して図6のように1~5文の質問文を作り、合計 243 文の質問文を用意した。情報文と質問文の 243 セットを目視により正解であるかを評価した。全体の評価と疑問詞別の評価を図7に示す。なお、図6の「○」は質問文に対して正しい回答と常識的に判断できるもので、「×」は質問文に対して明らかに無駄なものが含まれた回答、間違った回答、何も返ってこなかった回答を示している。

情報文：「794年、桓武天皇は京都に平安京を築いた。」
 質問文：「794年、桓武天皇はどこに平安京を築きましたか?」
 「桓武天皇が京都に平安京をいつ築いたのですか?」
 「794年、桓武天皇はどうしましたか?」

図6 情報文と質問文のセットの例

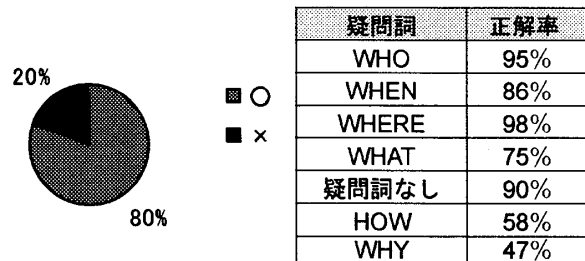


図7 評価結果

全体で80%の正解率を得ることができた。次に、成功例を図8に、失敗例を図9に示す。

情報文：「アメリカ大陸の発見者はコロンブスである。」
 質問文：「アメリカ大陸を発見したのは誰ですか?」

人物	用語	物	時間	場所	用言
コロンブス				アメリカ大陸	ある

質問対象語：人物
 条件：アメリカ大陸、発見
 疑問詞：WHO
 回答候補：コロンブス

図8 成功例

情報文：「喜田川歌麿は美人画で有名な人物である。」
 質問文：「喜田川歌麿はどのような人物ですか?」

人物	用語	物	時間	場所	用言
喜田川歌麿	美人画				ある
人物					

質問対象語：様子
 条件：喜田川歌麿、人物
 疑問詞：HOW
 回答候補：美人画人物である

図9 失敗例

間違えた原因としては以下のようなものが挙げられる。

1つ目は、意味理解システムを利用している部分での[成分]への誤った格納による間違いである。これは、意味理解システムに問題があるため、[成分]への格納方法を改善すると、精度が上がると考えられる。2つ目は、図9のような、知識フレーム内の語だけでは答えられなかったことによる間違いである。この例では「有名な」という語が知識フレームに格納されていないための問題だと考えられる。これに関しては、知識フレーム以外の情報を利用した新たな手法を考える必要がある。

5. おわりに

本稿では、与えられた情報文に対して、質問文が入力された場合に正しい回答を出力するといった、質問文を意味解釈する手法を提案した。質問文の疑問詞によっては、間違った回答が多く、正解率が低い疑問詞もあるため、処理の方法を考え直す必要がある。

今後は、情報文が知識文の単文で、文中に回答があるということを前提とせず、ある質問文が与えられたとき、数ある知識文の中から回答を含む文章を検索し、正確な回答を行えるように拡張させ、知識文だけでなく、日常使われるような会話文にも対応させていくことで、より人間に近い会話が可能になると考えられる。

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ものである。

参考文献

- [1] 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”, 人工知能学会全国大会, 2D1-10 (2004)
- [2] 篠原宜道, 渡部広一, 河岡司, “常識判断に基づく会話意味理解方式”, 言語処理学会第8回年次大会発表論文集, B6-2, pp.651-654 (2002)
- [3] 眞鍋康人, 小島一秀, 渡部広一, 河岡司, “概念間の関連度やシンソーラスを用いた概念ベースの自動精練手法”, 同志社大学理工学研究報告, Vol.42, No.1, pp.9-20, (2001)
- [4] 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.3, No.1, pp.53-74 (2006)