E-012

質問文の意味を考慮した Wikipedia からの回答文抽出手法
Method of extracting answer from Wikipedia considering meaning of question

森 泰宏† 芋野 美紗子† 土屋 誠司‡ 渡部 広一‡
Yasuhiro Mori Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

現代では工場で活躍している産業ロボットが主流であるが，社会の少子高齢化に伴い，今後は病院や福祉施設，家庭内といった場所での活躍が期待されている．そのためには，介護の手助けや円滑な会話を行うといった「日常生活の中で人間のパートナーとなる知的ロボット」の開発が必要である．人間は会話の中でよく質問をしたり，質問に答えるなどの受け答えを行うことでコミュニケーションをとる．その際，常識的な知識から連想や判断を行い，円滑な会話を実現させている．そのため，ロボットが人間と円滑な会話を行うには人間が持つ常識的な知識をロボットに持たせる必要があり，それにより与えられた知識に対する質問に答えることが可能になる．

常識には「朝の挨拶はおはよう」というような生活に関する知識や，「日本はアジアの島国である」というような教養に関する知識などがある．質問文に回答を行うには，予め質問文に関する知識文をロボットに与える必要があるが，教養知識の量は膨大であり，すべての知識をロボットに与えることは不可能である．そこで Wikipedia を用いることで質問に応じて必要な教養知識をロボットに与えることが可能となる．本稿では，質問文より連想を行い，Wikipedia から得られた教養知識の中で回答を含んでいる可能性のある文を抽出する手法を示す．

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連想することが可能である．この連想を模倣したものを連想メカニズムと呼ぶ．しかし，コンピュータは語と語の関連性を考慮することが難しく，人間のように連想を行うことは困難である．そこで連想メカニズムに必要な技術を以下に説明する．

2.1 概念ベース

概念ベース[1]は電子化された国語辞書や新聞記事などから，自動的に構築された知識ベースである．見出し語(概念)に対して，その特徴を表す語(属性)および属性の重要さ(重み)の対を複数付与することで構成されている．ある概念 A は m 個の属性 a_i と重み w_i (>0)の対によって次のように表現する．

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

2.2 関連度計算方式

関連度計算方式[2]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を計算し，その結果を数値で表現する．関連度は 0.0 以上 1.0 以下の連続的な実数であり，関連度が高いものほど関連の深い語であることを示す．

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

3. Wikipedia からの回答文抽出手法

質問文の入力から Wikipedia を用いて回答文を出力する流れを図 1 に示す．また，質問文「マラソンの正規走行距離は何キロメートルですか」(例文 1)を例に挙げて回答文の抽出手法を説明する．

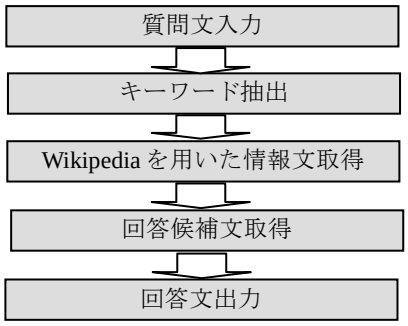

図 1 Wikipedia からの回答文抽出手法の流れ

3.1 キーワード抽出

入力される質問文から Wikipedia で検索する際に用いるキーワードを抽出する．そこで質問文に形態素解析を行い，名詞を抽出する．例文 1 の場合，キーワードは「マラソン，正規，走行，距離，何，キロメートル」となる．

3.2 Wikipedia を用いた情報文取得

質問文から抽出したキーワードに対して Wikipedia で検索を行う．そして，出力されたページでキーワードに関して説明している全文章を情報文として取得する．ここでの情報文とは複数の文の纏まりとする．

3.3 回答候補文取得

3.2.節で取得した情報文から回答候補文の取得を行う．回答候補文とは，情報文の中で，質問文に対する回答が含まれる可能性が高い一文のことである．回答候補文を取得する方法として表記一致検索と質問文意味理解システムを用いた手法を説明する．

3.3.1 表記一致検索

表記一致検索とは，情報文の中から質問文の名詞を含む文を抽出する処理である．質問文の名詞が一語でも含む文を全て抽出する．表記一致検索を行う理由は，質問文に対する回答文には，質問文の名詞が含まれている可能性が高いと考えられるからである．

3.3.2 質問文意味理解システムを用いた手法

質問文意味理解システム[3]とは，質問文から質問対象語と条件を取得するシステムである．質問対象語とは質問文の答えを大局的に表現する語のことであり，条件とは

質問対象語を修飾する語のことである．例文 1 の場合，質問対象語は「長さ」となる．

最初に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を取得する．3.3.1 節の表記一致検索によって，絞り込んだ情報文を X とする．次に，情報文 X の中で質問対象語と意味的に同じ語が含まれている文を抽出する．質問対象語と意味的に同じ語とは，質問対象語が「長さ」の場合，「キロメートル，メートル，センチメートル」といった長さに関する語のことを指す．質問対象語と意味的に同じ語の取得方法は，質問文意味理解システムに予め登録されてある単位データベースを参照している．単位データベースの例を表 1 に示す．

表 1　単位データベースの例

ID	単位	意味
99	センチメートル	長さ
127	キロメートル	長さ
142	平方メートル	面積

質問文意味理解システムの単位データベースには各単語に ID と意味が記載されている．表 1 より意味が「長さ」であるのは「センチメートル」と「キロメートル」であることから，質問対象語として取得する．「平方メートル」は「面積」であるため取得しない．

質問対象語と意味的に同じ語を取得した後，その取得した語を含んでいる文を情報文 X から抽出する．以上の方法で情報文から回答候補文を取得する．

3.4 回答文出力

回答候補文を取得後，質問文と回答候補文で関連度の算出を行う．回答候補文を「文 1，文 2，・・・，文 n」と仮定した場合，まず初めに質問文と文 1 で関連度の算出を行う．次に質問文と文 2 で関連度の算出を行う．同様に質問文と回答候補文の 1 文ずつ n 個の文に対して関連度の算出を行う．質問文の名詞に 3.3.2 節で得られた質問対象語を追加したものと，回答候補文一文に存在する名詞とを総当りで関連度の算出を行い，その値の平均値を質問文と回答候補文の関連度としている．例えば図 2 に示す例では質問文の名詞"マラソン"と文 1 の名詞"マラソン"の関連度は 1.0 となる．また質問対象語"長さ"と"マラソン"ならば 0.02 となり，同様に総当りで関連度を算出する．そして質問文との関連度が高い文から昇順で出力する．質問文の名詞に質問対象語を追加して関連度の算出を行う理由として，質問対象語と関連が深い文ほど，回答語が含まれている可能性が高いと考えられる．

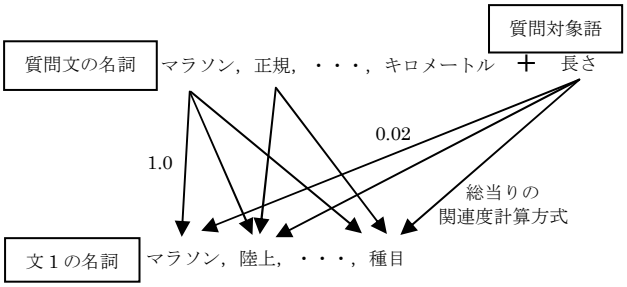

図 2　総当りによる関連度計算方式の例

4. 評価・考察

教養知識に関する質問文を 100 文用意し，提案手法の評価を行った．回答文として出力する文中に回答語が含まれる場合を正解とし精度を算出する．3 章の例文 1 の場合，「マラソンは陸上競技の長距離走のひとつで 42.195 キロメートルを走り，順位や時間を競う種目である．」という文が出力された．例文 1 の答えは「42.195 キロメートル」であるため，出力された文は正解となる．また回答語が含まれている文が関連度計算結果の上位 1 件に出力された場合，上位 2 件以内，上位 3 件以内，上位 4 件以内，上位 5 件以内に出力された場合で評価を行った．

結果として，上位 1 件に回答を含む文が出力されたのは 20%，上位 2 件以内で 30%，上位 3 件以内で 45%，上位 4 件以内で 54%，上位 5 件以内で 57%となった(図 3).

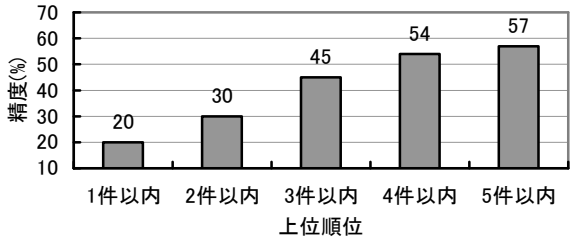

図 3　提案手法の評価結果

「野球は何人でするスポーツですか」という質問文に対し，質問文意味理解システムを用いなかった場合，回答となる語が含まれている文が関連度上位 8 件目に出力されたが，質問文意味理解システムを用いた場合，関連度上位 3 件目に出力された．その理由として関連度上位に出力されている文の中で，質問対象語と意味的に同じ語（質問対象語は人数であるため，意味的に同じ語は人や名である）を含んでいない文 5 件を雑音とて省いたことが挙げられる．

5. おわりに

本稿では教養知識に関する質問文に対し，質問文の意味を考慮し，Wikipedia を用いて回答となる文を抽出する手法を提案した．評価の結果，57%の精度で質問文の回答となる文を抽出することが可能となった．また，教養知識の量は膨大であり，すべてをロボットに与えることは困難であったが，Wikipedia を用いることで質問に応じて必要な教養知識をロボットに与えることが可能となった．

謝辞

本稿の一部は，科学研究費補助金（若手研究（Ｂ）24700215）の補助を受けて行った．

参考文献

[1]奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007.

[2]荻原 寛，渡部 広一，河岡 司，“概念ベース内の共起情報に着目した概念間関連度計算方式”，信学技報，Vol.106，No.587，pp.17-22，2007.

[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”，人工知能学会全国大会，2D1-10，2004.