D-011

Wikipedia を情報源とした質問応答システムの検討
Consideration of question answering system based on Wikipedia

相濱佑介† 土屋誠司‡ 渡部広一‡
Yusuke Aihama Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，ロボットが教育や福祉等の様々な分野で人間のパートナーとして活躍することが期待されている．そのためには人間と円滑にコミュニケーションをとるロボットが必要になる．人間は会話によりコミュニケーションをとるが，会話の中では質問応答が繰り返される．そのため，人間とロボットが円滑なコミュニケーションをとるためには質問応答が必要不可欠である．しかし，ロボットが人間のように質問応答を行うには，ロボットはある情報源から質問の回答を獲得する必要がある．そこで本稿では，Wikipedia を情報源として与えることで，Wikipedia から回答を獲得し，質問応答を行うシステムを提案する．

Wikipedia は大規模 Web 百科事典であり，幅広い分野について，一般的な事柄から新しい事柄に至るまで様々な記事が網羅されている．Wikipedia 日本語版には，2018 年 6 月 29 日時点で約 111 万もの膨大な数の記事が公開されており，Wikipedia のコンテンツなどのデータは，再配布や再利用のために利用できる一元化されたデータベース・ダンプでの提供が行われている．Nature 誌の調査[1]によると，Wikipedia の記事の精度は，専門家によって作成されたブリタニカ百科事典と同等であると報告している．よって，Wikipedia は信頼できる情報源であると考えられる．

なお，本稿では 2018 年 6 月 29 日時点の Wikipedia のデータを処理の対象にした．

2. 関連技術

2.1 概念ベース

概念ベースは電子化された国語辞書などから自動的に構築された知識ベースである．ある語を概念と定義し，概念の特徴を表す語である属性と，属性の重要さを表す重みの対で構成されている．

2.2 関連度計算

関連度計算方式[2]とは，概念ベースに定義されている 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の間の実数値で表され，概念間の関連が強いほど大きな値を示す．

2.3 質問文意味理解システム

質問文意味理解システム[3]は，係り受け解析器「南瓜」[4]を利用して質問文から質問対象語を取得するシステムである．質問文意味理解システムの使用例を表 1 に示す．

† 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

表 1　質問文意味理解システムの使用例

質問文	質問対象語
ノーベル物理学賞を受賞した日本人は誰ですか。	人物
世界で一番大きい鳥は？	鳥

質問対象語は質問文への回答の種類となる語で，表 1 の例だと，例えば「ノーベル物理学賞を受賞した日本人は誰ですか？」という質問に対して「人物」という語が質問対象語となる．質問文中に疑問詞「誰」や「場所」の表現があった場合質問対象語として「人物」，「場所」を獲得できる．また，疑問詞が「何」の場合や疑問詞がない場合でも質問対象語を獲得できる．

2.4 質問応答システムの概要

質問応答システムの構成図を図 1 に示す．このシステムは質問文解析，回答候補の獲得，回答スコア付けの 3 つの要素で構成される．

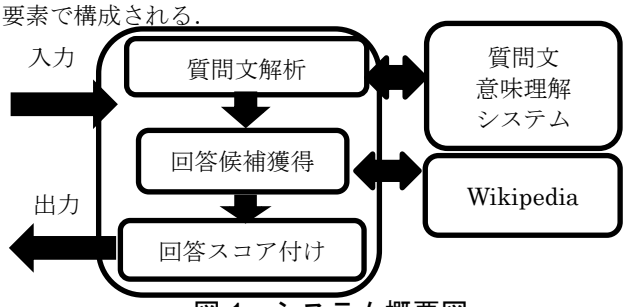

図 1　システム概要図

3. 質問文解析

質問文解析では質問文の解析を行い，質問対象語と検索語を獲得する．質問対象語の獲得については 3.1 節で，検索語の獲得については 3.2 節で説明する．

3.1 質問対象語の獲得

質問対象語獲得の流れを説明する．まず質問文意味理解システムに対して質問文を入力し，質問対象語を獲得する．得られた質問対象語が「場所」であった場合，関連度計算を用いて「場所」か「組織」のどちらかに質問対象語を変更する．これは，質問対象語「場所」は疑問詞「どこ」から導かれたと考えられるためである．疑問詞「どこ」から質問対象語を得た場合，場所を聞いている場合と組織を聞いている場合の両方が考えられる．例えば，「カレーに使われるスパイスのターメリックの原産国はどこか？」という質問文の場合は場所を聞いているが，「自動車「プリウス」を販売している企業はどこか？」という質問文の場合は組織について聞いている．そのため，質問対象語が「場所」となった場合は，「場所」，「組織」と検索語との間で関連度計算を行い，高い関連度を得られたほうを質問対象語としている．また，質問対象語が「名前」，「名称」のどれかであった場合，直前の名詞を質問対象語として獲

得する．(例：「商品の名前は」という部分がある場合，質問対象語は「商品」となる．)

上記以外の質問対象語の場合，それをそのまま質問対象語としてスコア付けに使用する．

3.2 検索語の獲得

検索語とは，質問文内の名詞，形容詞，動詞，副詞の集合である．質問文に対して MeCab[5]を用いて形態素解析を行い，検索語を獲得する．具体的な質問文と検索語の例を表 2 に示す．

表 2　検索語の獲得例

質問文	検索語
世界一大きい鳥は何か？	「世界一,大きい,鳥」
毎年 5 月にフランスで開かれている国際映画祭は何か？	「毎年,5 月,フランス,開く,国際映画祭」

4. 回答候補の獲得

獲得した回答候補に対して，回答スコア付けを行う．回答スコア付けの観点として，質問対象語に適するかどうか，質問文に適するかどうかの 2 つが存在する．回答候補が質問対象語に適するかどうかを調べるため．質問対象語と回答候補の関連度を使用し，回答候補が質問文に適するかを調べるため，検索語の名詞と回答候補の関連度を使用する．

最終的なスコアは質問対象語に適するかを考慮したスコアと質問文に適するかを考慮したスコアの積により求められる．

5. 評価手法

本稿で提案している手法の評価を行うために，テストセットとして，20 名にアンケートを行い収集した質問文 120 問を用いる．なお，テストセットの質問文には，正解となる回答を与えている．テストセットの例を表 3 に示す．

表 3　テストセットの例

質問文	正答
世界一大きい鳥は何か？	ダチョウ
毎年 5 月にフランスで開かれている国際映画祭は何か？	カンヌ国際映画祭

質問応答システムでは，質問文の質問対象語を正しく獲得できるかと回答候補の獲得時に正答が獲得できるかが回答を選択する上で重要になる．そこで，質問の質問対象語の獲得の評価，回答候補の獲得の評価を行った上で，回答スコア付けを含めた質問応答システム全体の評価を行う．

質問対象語の獲得の評価では正しい質問対象語を獲得したか，回答候補の獲得の評価では獲得した回答候補の中に正答となる回答候補が含まれているか，提案手法全体の評価ではスコア付けされて出力された回答の内，5 位以内に出力された回答が用意した正答と一致するかどうかで評価を行った．

6. 評価結果

テストセットの質問文 120 文を用いて評価を行った．質問対象語と回答候補の獲得の評価結果を表 4，質問応答システム全体の評価結果を表 5 に示す．

表 4　質問対象語と回答候補の獲得の評価結果

	質問対象語	回答候補の獲得
精度(%)	77	71

表 5　質問応答システム全体の評価結果

順位	精度(%)
1 位	11
2 位	7
3 位	5
4 位	6
5 位	2
正答なし	69

表 4 から，質問文の 77%に対して正しい質問対象語が得られ，質問文の 71%に対して正答となる回答候補が得られることが分かった．次に表 5 から，質問応答システム全体で上位 1 件に回答を含む文が出力されているのは 11%，上位 2 件では 7%，上位 3 件では 5%，上位 4 件では 6%，上位 5 件では 2%の結果が得られることが分かった．

7. 考察

結果として，質問文 120 問に対して出力 5 位以内に正答が存在する割合（31%）で正答を選択することに成功した．しかし，正答率は低く実用的な精度とは言い難いため，改善を行う必要がある．正答率が低くなった理由として，回答候補の獲得の精度が 71%と低いこと，スコア付けが適切でないことが挙げられる．

回答候補の獲得では，回答候補の獲得時に正答が獲得できるかは直接的に精度に関係してくるため，今回用いた獲得法では 71%と低いため，質応答システム全体の精度は低くなったのではないかと考える．また，スコア付けでは質問対象語と回答候補の関連が強い場合でも，関連度が低く出力される場合が幾つか見られた．このため，質問応答システム全体の精度が低くなったのではないかと考える．

8. おわりに

本稿では，Wikipedia を情報源とした質問応答システムを提案した．手法として，語が有する意味特徴を語と重みで表現する概念ベースと語と語の間にある意味的な関連性を数値として算出する関連度計算方式，質問文意味理解システム，Wikipedia 日本語版のダンプデータを利用した．

今後の課題としては,回答候補の獲得の精度向上,回答スコア付けの見直しが挙られる．これらを改善することで質問応答システムの精度向上ができるのではないかと考える．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けて行ったものです．

参考文献

[1] Giles, J.,“Internet Encyc1opaedias Go Head to Head”, Nature, Vol.438, pp.900-901, 2005.

[2] 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，pp.159‐160，2002.

[3] 奥村紀之, 荒木孝允, 渡部広一, 河岡司，“概念属性の動的評価に基づく概念関連度計算方式”，情報処理学会，E-033，pp.223-226，2006.

[4]“形態素解析システム「南瓜」”，松本雄二

[5]“MeCab--態素解析器”，http://taku910.github.io/mecab/，京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所，2019-2-11 参照．