回答の根拠を提示する意思決定支援型の質問応答システム

佐々木 智† 藤井 敦‡
† 筑波大学大学院図書館情報メディア研究科
‡ 東京工業大学大学院情報理工学研究科

1 はじめに

インターネットの普及に伴い、多種多様な情報がWebに発信されるようになった。大量のWeb文書から、ユーザの欲する情報を効率良く見つける手法として質問応答(QA)がある。

QAは人工知能と自然言語処理の分野で研究されている。前者はシステム固有の形式で組織化された情報を用いて回答を推論する「推論型」である。後者は組織化されていない文書集合から回答を抽出する「抽出型」である。推論型QAは情報の組織化が高価であり、拡張性が乏しく回答できる分野が限定される。そのため、近年では抽出型QAに関する研究が活発である。

抽出型QAは、対象とする質問の種類によって手法を分類することができる。名称、日付、数値など客観的事実を問う質問に回答するQAは「factoid型」、行動、原因、定義などを問う質問に回答するQAは「non-factoid型」と呼ばれる。non-factoid型は、質問の種類により、行動や手順を問う質問に回答する「how型」、原因や根拠を問う質問に回答する「why型」などに分かれる。単一の手法でnon-factoid型に属する全種類の質問に回答する手法が提案されている[1]。しかし、この手法は大規模なFAQコーパスを必要とする。本研究では、how型QAに焦点を当てて探求する。

how型QAの研究事例として、ヘルプデスク型QA[2]がある。このQAは、述語と項の対を用いて行動を問う質問に回答するため、意思決定を支援するシステムと見なすことができる。例えば、蜂に刺された時の対処法について意思決定をしたいユーザがいるとする。ユーザはヘルプデスクQAに「蜂に刺されたらどうすればいい?」という質問を入力することで、「針を抜く」、「救急車を呼ぶ」など取るべき行動の選択肢を得ることができる。

しかし、意思決定をする上で、どの行動が適切か判断する必要がある。そのための支援として、本QAシステムは各行動表現と共にその行動を取るべき理由を提示する。例えば、「救急車を呼ぶ」という行動表現に対して「蜂毒アレルギーのある人は、一刻も早く医者の診断が必要なので」という理由が得られた場合、ユーザが蜂毒アレルギーを持つかどうかで「救急車を呼ぶ」という行動が適切かどうか判断することができる。

how型QAの手法は、ヘルプデスク型QA[2]の他、Moriら[1]や渡辺ら[3]も提案している。しかし、いずれの手法も取るべき行動を回答するだけであり、その理由も合わせて答える手法はない。QAで出力された回答に対して理由を抽出する手法は、factoid型QAが対象である[4]。

以上の背景を踏まえ、我々は取るべき行動を理由と共に答えるQAシステムを提案した[5]。以降では、本QAシステムの構成と既存の情報検索手法と比較評価した結果について順番に説明する。

2 本QAシステムの構成

2.1 概要

図1に基づいてシステムの動作について説明する。ユーザは、「蜂に刺されたらどうすればいい?」といった行動を問う質問文を入力する。「how型QA」は質問に対する回答として、「患部を洗う」や「アウトドアに行く」といった行動表現を出力する。ここで、前者は正解であり、後者は誤答である。さらに、行動表現を含む文章を記述的な回答として出力する。「理由を問う質問文の生成」は、how型QAに入力された質問と出力された回答を用いて、「なぜ蜂に刺されたら患部を洗うのか?」といった理由を問うwhy型質問文を生成する。この質問を「why型QA」の入力とし、「患部を洗う」という行動を取るべき理由を回答する。「回答の統合」はhow型QAで得られた回答とwhy型QAの回答を組み合わせて一つの回答にする。回答のスコアを再計算し、理由が抽出されなかった行動表現は順位を下げる。

上記の例では、「患部を洗う」という行動表現は蜂に刺された時の対処法として正しいため、why型QAで理由が抽出される可能性が高い。しかし、「アウトドアに行く」は、蜂に刺された時の対処法として不適切であるため、why型QAで理由が得られない。そこで、「患部を洗う」を「アウトドアに行く」よりも上位にする。その結果、how型QAを単体で使うよりもwhy型QAと組み合わせることでシステムの精度を高めることができる。

2.2〜2.5節で、how型QA、理由を問う質問文の生成、why型QA、回答の統合についてそれぞれ説明する。

2.2 how型QA

図1のhow型QAには、ヘルプデスク型QA[2]を拡張して用いる。本how型QAは、入力された質問文と

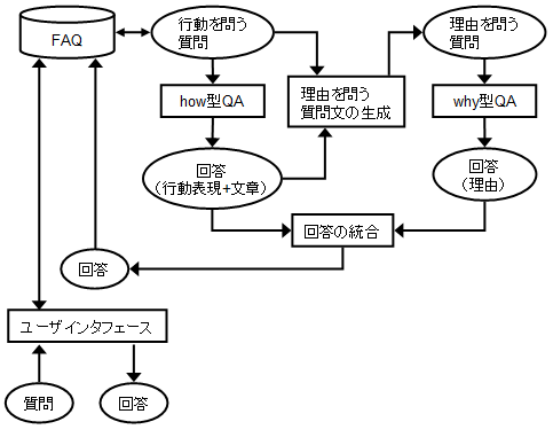

図 1: 本研究で提案する質問応答システムの構成

関連のある文書を Web から収集する「情報検索」、収集された文書から行動表現を抽出する「回答抽出」、適切な行動表現及びその行動表現が含まれる文章により高いスコアを付ける「回答評価」の順に処理を行う。

「回答抽出」では、収集された文書を係り受け解析し、述語と項の対を全て抽出して、行動表現の候補とする。ただし、以下の条件を満たす述語と項の対は抽出しない。

- **一般的な表現である。**
  「気がする」のように、名詞「気」や動詞「ある」、「する」、「なる」、「やる」を含む表現は誤答であることが多い。
- **Web に頻出する表現である。**
  「トップページに戻る」などの Web に頻出する表現は、誤答であるにも拘らず回答候補として抽出されやすい。そこで、Web に頻出する表現のリストを人手で作り、リストに登録されている表現は抽出対象から削除する。
- **質問文に含まれる表現である。**
  「ニキビができたらどうすればいい?」という質問に対して、「ニキビができる」という表現は回答として不適切である。

「回答評価」における行動表現のスコア付けでは、以下に示す a〜e の基準に合致する行動表現に高いスコアを与える。

a **名詞句 (名詞+助詞) と動詞の係り受け距離が近い。**
係り受けの距離とは、名詞句と動詞の間にある形態素数である。この距離が短いほど、その名詞句と動詞の関連は強いと考える。また、距離が短いほど一般的に係り受け解析の誤りが少ないため、係り受け関係にあることの確実性が高い。

b **推奨表現や禁止表現を伴う。**
推奨表現（「〜すること」や「〜しましょう」など）は問題解決に有効な対処法を述べる時に用いられる。禁止表現（「〜してはいけない」など）は行ってはならない対処法を述べる時に用いられ、推奨表現と同様に有用である。

c **抽出元ページの検索結果における順位が高い。**
行動表現が抽出されたページの順位が高いほどスコアを上げる。具体的には式 (1) を用いる。

$$\frac{\text{検索ページ数} - \text{抽出元ページの順位}}{\text{検索ページ数}} \quad (1)$$

d **質問に含まれる行動表現との距離が近い。**
距離とは、行動表現中の動詞と質問中の動詞の間にある形態素数である。この距離が短いほど、その行動表現は質問に対して強い関連性を持つと考える。

e **抽出元の文でガ格が係らない。**
例えば「蜂が巣を守る。」という文から「巣を守る」という行動表現が得られる。しかし、抽出元の文で「蜂が」というガ格の名詞句が係り、動作主は「蜂」と分かる。ゆえに、質問者がすべき行動として不適切である。

基準 a〜e を式 (2) によって統合し、行動表現 x のスコア $s(x)$ を計算する。

$$s(x) = \sum_i \left(\frac{1}{a(x_i)} + b(x_i) + c(x_i) + \frac{1}{d(x_i)}\right) \cdot e(x_i) \quad (2)$$

検索された複数の Web 文書において、同じ行動表現が繰り返し出現することがあるため、i 番目に出現する x のスコアをそれぞれ求め、それらの総和を x のスコアとする。$a(x_i)$ は係り受けの距離である。$b(x_i)$ は x_i が推奨・禁止表現を伴なえば 1 であり、伴わない場合は 0 である。$c(x_i)$ は式 (1) で計算する。$d(x_i)$ は質問との距離である。$e(x_i)$ は x_i にガ格の名詞句が係らなければ 1 であり、係る場合は 0 である。

更に、$s(x)$ を情報検索の重み付け手法である IDF によって補強する。IDF は、「情報を集める」や「他人に聞く」など、多くの質問に共通して出現しやすく特定の質問と関連しない行動表現に対してスコアを下げる効果がある。

$$s_{idf}(x) = s(x) \cdot IDF(x) \quad (3)$$

文章 p のスコア $s(p)$ は、その文章に含まれる行動表現のスコアを総和して求める。

$$s(p) = \sum_{x \in p} s(x) \quad (4)$$

2.3 理由を問う質問文の生成

how 型 QA で得られた行動表現に対し、その行動を取るべき理由を問う why 型質問文を生成する。生成は以下の手順に沿って行う。

1. 行動を問う質問文から「動詞句+たら」または「動詞句+には」という記述を抽出する。例えば、「蜂に刺されたらどうすればいい?」という質問文からは、「蜂に刺されたら」という記述が抽出される。

2. 1. で抽出された記述と how 型 QA で回答として得られた行動表現を連結する。例えば、「蜂に刺されたら」と「患部を洗う」からは、「蜂に刺されたら患部を洗う」が得られる。
3. 2. で得られた記述に「なぜ」と「のか?」を付けて質問文とする。例えば、「蜂に刺されたら患部を洗う」という記述からは、「なぜ蜂に刺されたら患部を洗うのか?」という質問文が生成される。

2.4 why 型 QA

図 1 の why 型 QA には、渋沢ら [6] の手法を拡張して用いる。本 why 型 QA も how 型 QA と同様に、「情報検索」、「回答抽出」、「回答評価」の順に処理を行う。渋沢らは、why 型質問の内容を表す文を「質問相当文」と定義した。「情報検索」において収集された文書から質問相当文を探し、その周辺にある手掛かり表現を伴う文を回答候補として抽出する。手掛かり表現には、「なので」や「ゆえに」など理由の記述に特有の表現を人手で定義して用いる。

how 型 QA の「情報検索」において検索された Web 文書にも、why 型 QA の回答が含まれている可能性がある。しかし、how 型 QA における Web 検索は why 型 QA の回答が含まれる文書の収集を目的としていない。そこで、検索の手間よりも無関係な文書を減らすことを重視し、how 型 QA で検索された文書は why 型 QA の回答抽出に用いず、why 型 QA の「情報検索」において再度 Web を検索する。

「回答評価」では、行動表現のスコア付けに用いた基準 c と以下に示す f〜i に合致する回答候補に高いスコアを与える。

f **重みの大きい理由語を多く含む。**

「なので」や「ゆえに」などの理由の記述に特有の表現を多く含むほど、その文は理由の記述である可能性が高い。また、「なぜかというと」などの明らかに理由の記述に出現する表現は重みを大きくし、「故」や「理由」などの理由を表さない記述にも出現する表現は重みを小さくする。

g **質問相当文との距離が近い。**

抽出元ページにおいて j 番目の文が質問相当文で k 番目の文が回答候補である場合、回答候補と質問相当文の距離を式 (5) によって計算する。この距離が短いほど、その回答候補は why 型 QA に入力された質問の内容と強い関連性を持つと考える。

$$|k - j| \tag{5}$$

h **回答候補や前後に how 型質問中の単語が出現する。**

「2.3 理由を問う質問文の生成」で用いられた how 型質問文に含まれる名詞または動詞が回答候補や前後の文に出現する場合、その回答候補は how 型質問の内容と関連性があると考える。例えば、「蜂に刺されたらどうすればいい?」という how 型質問文が用いられた場合、「蜂」や「刺す」という単語が近くに出現する回答候補ほど、その内容は蜂に刺されたことに関する記述である可能性が高い。

i **質問相当文に含まれる行動表現にガ格が係らない。**

例えば、「蜂に刺されたらどうすればいい?」という how 型質問と「患部を洗う」という行動表現を基に、「血がにじみ出たため、蜂に刺された患部を血が洗ってくれた。」という質問相当文が得られたとする。この質問相当文において、「患部を洗う」という行動表現には「血が」というガ格の名詞句が係り、「患部を洗う」という行動の動作主は質問者ではなく「血」と分かる。ゆえに、この質問相当文からは、質問者が「患部を洗う」という行動をすべき理由が得られない可能性が高い。

基準 c と f〜i を式 (6) によって統合し、回答候補 y のスコア $s(y)$ を計算する。

$$s(y) = c(y) \times f(y) \times (L - g(y)) \times h(y) \times i(y) \tag{6}$$

$c(y)$ は式 (1) で計算する。$f(y)$ は y に含まれる理由語が持つ重みの総和である。$g(y)$ は式 (5) で計算する。$h(y)$ と $i(y)$ は 0 か 1 の値を取る。$h(y)$ は質問相当文に含まれる行動表現にガ格が伴えば 0、$i(y)$ は y または y の前後文に how 型 QA で入力された質問に含まれる単語が出現すれば 1 である。$f(y)$ と $g(y)$ は渋沢ら [6] が提案するスコア付けの基準に基づいており、$c(y)$、$h(y)$、$i(y)$ は我々が提案した基準 [5] である。

2.5 回答の統合

how 型 QA で得られた行動表現と、各行動表現に対して why 型 QA で得られた理由を一つの回答として統合する。行動表現に与えられたスコアと理由に与えられたスコアも統合する。その結果、理由が伴わない不適切な行動表現は順位が下がる。

式 (7) を用いて行動表現と理由のスコアを統合する。$s'(x)$ は行動表現 x のスコア、$s'(y)$ は行動表現 x に対して得られた理由 y のスコアを正規化した値である。正規化にはシグモイド関数を用い、$s'(x)$ と $s'(y)$ の取り得る値の範囲が等しくなるようにした。

$$s(x, y) = s'(x) \cdot s'(y) \tag{7}$$

3 評価実験

評価実験では、how 型 QA に焦点を当てて評価を行った。評価には、「蜂に刺されたら」や「やけどをしたら」など、30 件の質問を用いた。各質問文をクエリとして Yahoo! JAPAN で Web 検索を行い、それぞれ上位 100 件のスニペット及び Web ページを収集した。正解判定はスニペット及び Web ページに対して行い、質問に対して正解の情報を含んでいるかどうかの 2 値判定とした。すなわち、QA を「スニペット及び Web ページの順位付け問題」とすることで、既存の情報検索手法との比較を可能にした。具体的には、以下に示す手法 A〜D を精度、再現率、F 値で比較した。

表 1: 上位 10 件の回答に対する評価結果
（各手法において、左右の数値はそれぞれスニペットと Web ページを回答単位とした場合の値を示す）

手法	A		B		C		D	
精度	0.620	0.833	0.597	0.780	0.777	0.783	0.683	0.883
再現率	0.116	0.124	0.113	0.115	0.139	0.109	0.131	0.140
F 値	0.187	0.210	0.181	0.196	0.234	0.191	0.210	0.234

表 2: 上位 10 件の回答に対する両側 t 検定の結果
（手法 C と D において、左右の表記はそれぞれスニペットと Web ページを回答単位とした場合の結果を示す）

評価尺度	手法 C		手法 D	
手法 A の精度	○	×	×	△
手法 A の再現率	×	×	×	△
手法 A の F 値	○	×	×	△
手法 B の精度	○	×	△	○
手法 B の再現率	×	×	△	△
手法 B の F 値	○	×	△	○

○:有意水準 1%で有意差あり、△:有意水準 5%で有意差あり、×:有意差なし

A : Yahoo! JAPAN
B : 単語の TF.IDF + PRF
C : 行動表現の 式 (2) + 式 (4)
D : 単語の IDF と行動表現の式 (3) + PRF

手法 B〜D の「X+Y」という表記において、X はタームのスコア付け手法、Y はスニペット及び Web ページのスコア付け手法を表す。A と B は従来の情報検索手法、C と D が本研究で用いている how 型 QA である。

本 how 型 QA は、タームの種類、タームのスコア付け手法、文章のスコア付け手法に何を用いるか幾つかの選択肢がある。全ての組み合わせを比較した結果、手法 C はスニペット、手法 D は Web ページを回答単位とした場合に最も高い F 値を示した手法であった。なお、手法 C と D は 2.2 節で説明した行動表現のスコア付け基準において、c と e のみを用いている。

評価結果を表 1 及び表 2 に示す。それぞれ、上位 10 件のスニペット及び Web ページにおける精度、再現率、F 値の比較である。質問 1 件では精度が良ければ再現率も必ず良くなる。しかし、30 件の質問において得られた値の平均であるため、精度で良い結果を示した場合に再現率でも良い結果を示すとは限らない。

スニペットを回答単位とした場合に最も高い F 値を示した手法である C を、既存の情報検索手法である A 及び B と比較する。表 1 において、スニペットを回答単位とした場合の値を比較すると、手法 C は A 及び B よりも全評価尺度において高い値を示した。表 2 において、有意水準 1%で精度と F 値に有意差が示された。

同様に、Web ページを回答単位とした場合に最も高い F 値を示した手法である D を、既存の情報検索手法である A 及び B と比較する。表 1 において、Web ページを回答単位とした場合の値を比較すると、手法 D は A 及び B よりも全評価尺度において高い値を示した。表 2 において、手法 D は手法 A に対し有意水準 5%で全ての評価尺度において有意差が示された。手法 D と手法 B を比較した場合においては、有意水準 1%で精度と F 値に有意差が示された。

以上より、行動表現をタームとして用いた手法は、回答単位としてスニペットまたは Web ページのどちらを用いても、既存の情報検索手法より良い結果を示した。すなわち、行動表現をタームとして用いる効果が示された。今後の課題として、タームの種類、タームのスコア付け手法、文章のスコア付け手法として考えられる各手法の有効性を確認する必要がある。

4 おわりに

本研究では、回答の根拠を提示する意思決定支援型の QA システムを提案した。評価実験では、既存の情報検索手法と比較して本 how 型 QA の有効性を確認した。今後は、評価実験に用いる質問数を増やし、how 型 QA と why 型 QA の両方を大規模に評価する必要がある。また、「国の借金を返すには」のように、何が正解であるかについて議論の余地がある質問への対応も今後の課題である。

謝辞

本研究の一部は、文部科学省科研費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（課題番号:21013003）によって実施された。

参考文献

[1] Tatsunori Mori, Takuya Okubo, and Madoka Ishioroshi. A QA system that can answer any class of Japanese non-factoid questions and its application to CCLQA EN-JA task. *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 41–48, 2008.

[2] 三原英理, 藤井敦, 石川徹也. Web を用いたヘルプデスク指向の質問応答システム. 言語処理学会第 11 回年次大会発表論文集, pp. 1096–1099, 2005.

[3] 渡辺靖彦, 西村涼, 岡田至弘. メーリングリストを利用した質問応答システムのための知識の内容確認. 情報処理学会研究報告, 2006-NL-174, pp. 55–59, 2006.

[4] Alvaro Rodrigo, Anselmo Perias, and Felisa Verdejo. Overview of the answer validation exercise 2008. *Working notes for the CLEF 2008 Workshop*, 2008.

[5] 佐々木智, 藤井敦. 取るべき行動を理由と共に答える質問応答システム -how 型と why 型の統合-. 言語処理学会第 15 回年次大会 発表論文集, pp. 36–39, 2009.

[6] 渋沢潮, 林貴宏, 尾内理紀夫. Why 型質問の回答文を Web ページから抽出するシステム RE:Why の試作. コンピュータソフトウェア, Vol. 24, No. 3, pp. 20–28, 2007.