取るべき行動を回答するhow型質問応答システムの評価

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）が注目されている。

QA は主に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も提案されている [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、三原らが提案した「ヘルプデスク型 QA」[2] がある。このシステムは、述語項構造を用いて行動を問う質問に回答する。例えば、「蜂に刺されたらどうすればいい?」という質問には、「針を抜く」や「患部を洗う」などを出力する。本研究では、述語項構造で表現される how 型 QA の回答候補を「行動表現」と定義する。また、「患部を洗う」という行動表現のみを回答として提示しても、どのような手順を取ればいいか詳細な内容がユーザには分からない。そこで、「患部を流水で洗い、毒を洗い流してください。それでも、気分が悪ければ、病院に行ってください。」のように、その行動表現を含んだ文章を記述的な回答として出力する。三原らの how 型 QA は、ユーザが抱えている問題に対して解決策を提示するため、ヘルプデスクやコールセンターの自動化を指向している。

筆者らは当該 how 型 QA に why 型 QA を統合し、回答の根拠を提示する QA システムを提案した [3]。本研究では how 型 QA に焦点を当てて評価を行った。

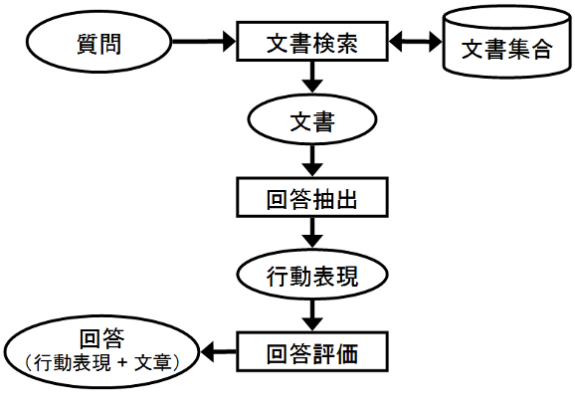

図 1: how 型 QA システムの構成

2 節で本システムの構成について説明し、3 節で評価実験について説明する。

2 how 型 QA システムの構成

2.1 概要

本研究で対象とする how 型 QA システムの構成を図 1 に示す。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。「文書検索」は入力された質問文をクエリに Web を検索し、質問文と関連のある文書を収集する。「回答抽出」は収集された文書から「針を抜く」や「患部を洗う」などの行動表現を抽出する。しかし、ここで抽出した表現の中には、「毒が回る」や「意識を失う」といった不適切な表現も含まれる可能性がある。そこで、「回答評価」では、不適切な表現に対して低いスコアが付くように、行動表現に対してスコアを付ける。また、行動表現のスコアに基づいて文章にもスコアを付ける。最終的に、端的な回答である行動表現と、記述的な回答である文章を出力する。2.2〜2.4 節で各構成要素について詳しく説明する。

2.2 文書検索

本研究では、文書検索で用いる文書集合として Web を利用する。ただし、原理的には、説明的な文書であれば、新聞記事など様々な種類の文書を利用することが

できる。Web には、Yahoo!知恵袋[1]を始めとする様々な FAQ サイトがあり、QA システムの情報源として有効である。しかし、企業や個人のホームページ、ブログにも how 型質問に対する回答は含まれている。

初期検索には、Yahoo! JAPAN[2]を用いる。Yahoo! JAPAN は入力された質問文を解析し、その質問文に含まれる単語やフレーズを文字列として含んでいるページを検索結果として返す。

2.3 回答抽出

原理的に、行動表現は「患部を流水で洗う」のように、複数の名詞句（名詞+助詞）と動詞から構成される述語項構造で表現する。しかし、複数の名詞句を考慮すると、「流水で洗う」と「患部を流水で洗う」が異なる行動表現で扱われるという問題が生じる。そこで、「患部を流水で洗う」という行動表現からは「患部を洗う」と「流水で洗う」の 2 つの行動表現に分けるように、1 つの名詞句と動詞で構成される述語項構造を行動表現の抽出単位とする。

検索された各 Web 文書に対して CaboCha[3]を用いて係り受け解析を行い、係り受け関係にある名詞句と動詞の組を行動表現として抽出する。ただし、以下に示す条件のいずれかを満たす表現は行動表現として抽出しない。

- **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。このように、質問文に含まれる行動表現は抽出しない。

2.4 回答評価

回答評価では、行動表現と、その行動表現が含まれている文章に対してスコアを付ける。行動表現のスコア付けでは、以下に示す a〜d の基準を用いる。

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、真に係り受け関係にあることの確実性が高い。

b **「〜すること」や「〜しましょう」のような推奨表現や「〜してはいけない」のような禁止表現である。**
推奨表現は問題解決に有効な対処法を述べる時に用いられる。禁止表現は行ってはならない対処法に用いられ、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、具体的に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。

基準 a〜d を満たす行動表現ほどスコアを高くする。具体的には、基準 a〜d 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。

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

検索された複数の Web 文書において、同じ行動表現が何回も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) で計算される値である。$d(x_i)$ は質問との距離である。

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。

$$s(p) = \sum_{x \in p} s(x) \quad (3)$$

本研究では、行動表現と文書のスコア付けそれぞれに対して修正を行った。行動表現のスコア付けでは、式 (2) で使用している a〜d の基準に加えて、情報検索の重み付け手法である IDF を用いる。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ユーザの情報要求を満たすには不十分な行動表現に対してスコアを下げる効果がある。一方、TF は修正に用いなかった。これは、how 型 QA の回答として適切な行動表現が同一の文書に何回も出現することは少ないためである。

式 (4) により、式 (2) において得られた行動表現 x のスコアを IDF で補強する。IDF は、式 (5) で計算する。式 (5) において、N は文書集合中の総文書数である。ただし、Web 全体の文書数は分からないので、N として十分に大きな定数を与える。

[1] http://chiebukuro.yahoo.co.jp/
[2] http://developer.yahoo.co.jp/webapi/search/
[3] http://chasen.org/~taku/software/cabocha/

$$s'(x) = s(x) \cdot IDF(x) \quad (4)$$

$$IDF(x) = \log \frac{N}{DF(x)} + 1 \quad (5)$$

文章に対するスコア付けでは、擬似フィードバック（PRF）[4] を応用する。具体的には、単語の代わりに行動表現を素性として用い、質問文と回答候補となる文章との類似度をその文章のスコアとして計算する。

3 評価実験

3.1 実験方法

評価実験では、本研究の修正によって、how 型 QA の有効性がどのように変化するかを評価した。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。各質問文をクエリに Yahoo! JAPAN で Web 検索を行い、それぞれ上位 100 件のスニペットを収集した。そのスニペットから行動表現を抽出し、スニペットをその行動表現を含む記述的な回答候補とした。正解判定はスニペットに対して行い、質問に対して正解の情報を含んでいるかどうかの 2 値判定をした。

評価尺度には、精度、再現率、F 値、MRR を用いた。また、誤答のスニペットよりも正答のスニペットに高いスコアを付けることができたか考察するために、正答と誤答のスニペットに付けられたスコアの平均と、その比率を計算した。

これらの評価尺度を用いて、以下の手法 A〜F を比較した。B〜F は、「素性のスコア付け手法 + 文章のスコア付け手法」という形式で使用した手法を示している。B は単語を素性、C〜F は行動表現を素性に用いた。

A : Yahoo! JAPAN
B : 単語の TF.IDF + PRF
C : 式 (2) + PRF
D : 式 (4) + PRF
E : 式 (2) + 式 (3)
F : 式 (4) + 式 (3)

A と B は従来の情報検索手法、E が三原らが提案した how 型 QA である。C、D、F は、本研究で提案した E の修正手法である。

素性の重みを計算する上で、単語の DF は、その単語をクエリに Yahoo! JAPAN で Web 検索を行い、その結果ヒットした Web ページ数とした。行動表現の DF は、TSUBAKI[4] を用いて収集した。TSUBAKI はクエリ中の係り受け関係を考慮しており、「患部を洗う」という行動表現で検索した場合、「患部を流水で洗う」という文字列を含んだ Web ページも収集することができる。

[4] http://tsubaki.ixnlp.nii.ac.jp/

表 1: 上位 10 件における各評価値の比較

手法名	A	B	C	D	E	F
精度	0.397	0.427	0.437	0.460	0.530	0.527
再現率	0.098	0.130	0.138	0.137	0.148	0.160
F 値	0.157	0.199	0.214	0.211	0.231	0.245
MRR	0.595	0.593	0.666	0.594	0.727	0.671

表 2: 正答と誤答に付けられたスコアの平均とその比率

手法名	B	C	D	E	F
正答の平均スコア	56.18	0.224	20.50	99.24	968.4
誤答の平均スコア	50.11	0.117	28.90	91.25	884.7
平均スコアの比率	1.121	1.916	0.709	1.088	1.095

※ 平均スコアの比率 = 正答の平均スコア / 誤答の平均スコア

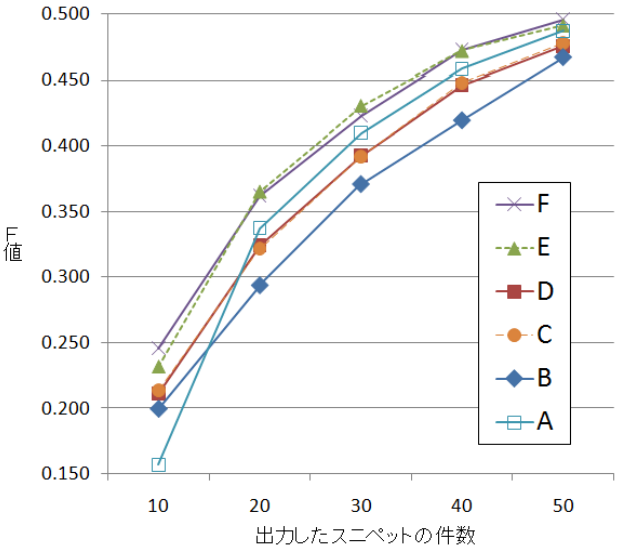

図 2: 異なるスコア付け手法の F 値

3.2 行動表現のスコア付けに関する評価

行動表現に対するスコア付け手法を比較した。1 つは三原らが提案した式 (2) によるスコア付け手法であり、もう 1 つは本研究で提案した式 (4) によるスコア付け手法である。C と E は式 (2) を用いており、それぞれの修正手法として D と F は式 (4) を用いている。

C と D を比較すると、表 1 では精度のみ D の方が高い値を示した。図 2 では、C と D はほぼ重なり、明確な差はなかった。表 2 では、C は正答の方が誤答よりもスコアの平均が高いのに対し、D は誤答の方が正答よりもスコアの平均が高く、式 (4) の効果は見られなかった。

E と F を比較すると、表 1 では再現率と F 値において F の方が高い値を示した。図 2 では、出力したスニペットの件数が上位 20〜30 件においては E の方が高い値を示したものの、それ以外の件数では F の方が高い値を示した。表 2 では、F の方が平均の比率がより高い値を示し、式 (4) の効果が見られた。

以上より、本研究の修正手法として式 (4) の有効性が

部分的に示された。

3.3 文章のスコア付けに関する評価

文章に対するスコア付け手法を比較した。1 つは行動表現を素性とする PRF である。もう 1 つは、三原らが提案した式 (3) によるスコア付け手法である。C と D は PRF を用いた手法であり、それぞれを三原らのスコア付け手法に置き換えた手法が E と F である。

表 1 において、C と E を比較すると E の方が、D と F を比較すると F の方が、全ての評価尺度においてより高い値を示した。図 2 においても、C と E を比較すると E の方が、D と F を比較すると F の方が、全体的に高い F 値を示した。表 2 では、平均の比率を比較すると、C と E では C の方がより高い値を示し、PRF を用いた効果が見られた。D と F では F の方がより高い値を示し、PRF を用いた効果は見られなかった。

以上より、PRF を用いることで、誤答のスニペットよりも正答のスニペットに高いスコアが付くよう修正した効果が部分的に見られた。その他の評価尺度においては、PRF は用いず、式 (3) を用いた方が有効であった。

3.4 考察

手法 C〜F において順位の低い正解のスニペットを分析した結果、以下の傾向があった。

- 行動表現として抽出したい表現が「名詞+助詞+動詞」という述語項構造になっていない。例えば、「やけどをしたら」という質問において得られた正解のスニペットに「充分冷やす」という記述があった。この記述は行動表現として抽出されるべきである。
- 「する」という動詞を含んでいる。例えば、「やけどをしたら」という質問において得られた正解のスニペットに「包帯をする」という記述があった。この記述は「名詞+助詞+動詞」という述語項構造をしている。しかし、「する」という動詞を含んでいる表現は抽出対象から外したため、抽出することができなかった。
- 手法 E および F においては、行動表現を多く含む文章ほど上位になる傾向がある。適切な行動表現を少しだけ含んでいるスニペットよりも、不適切な行動表現を多く含むスニペットの方が上位になる場合があった。

また、検索されたスニペットには類似した複数の文章が含まれていた。例えば、手法 E において「蜂に刺されたら」という質問の回答として上位 1 位と 2 位になったスニペットを図 3 に示す。類似したスニペットは含んでいる行動表現の種類と数が共通するため、同じスコアが付きやすい。結果として、「事故にあったら」という質問においては上位 40 件まで類似したスニペットが集中し、このスニペットが不正解のため、極端に精度や再現率が低くなった。

今後の課題として、「名詞+助詞+動詞」以外で表現される行動表現の構造を調査し、抽出対象に加える必要がある。行動表現の数が文章のスコア付けに大きく影響する問題については、不適切な行動表現に比べて適切な行動表現により高い重みを付けることで解決することができる。解決方法の一つとして、「針を抜く」と「針を取り除く」のように意味的に類似する行動表現を同じ行動表現として扱えるようにする点がある。また、図 3 のように類似した文章を 1 つにまとめる必要がある。

3 .. 毒液は水に溶けやすいため、蜂に刺されたら、まず傷口を流水でよく洗い流し、傷口から毒を絞り出しましょう。... 蜂に刺されたら、アンモニア水やおしっこをかけるとよいという考えもあるようですが、蜂の毒はアンモニアで中和するというのは、効果がありません。...
毒液は水に溶けやすいため、蜂に刺されたら、まず傷口を流水でよく洗い流し、傷口から毒を絞り出そう。 ... 蜂に刺されたら、アンモニア水やおしっこをかけるとよいという考えもあるようだが、蜂の毒はほとんど中性に近く、アンモニアで中和するというのは、 ...

図 3: 類似するスニペットの例

4 おわりに

本研究では、従来の how 型 QA に対して修正と評価を行った。具体的には、how 型 QA の回答単位となる行動表現と文章に対し、IDF と PRF をそれぞれ応用した。実験により、どちらの手法も部分的に有効性が示された。今後の課題は、質問の件数を増やして、より大規模な評価実験を行うことである。

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] 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.

[4] J.J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System - Experiments in Automatic Document Processing*, pp. 313–323. Prentice Hall, Inc., 1971.