

How-to 型質問応答システムにおける回答の選択手法

杉山 大輔*

* 東京理科大学大学院

太原 育夫†

† 東京理科大学

1. はじめに

現在、インターネットの普及により、Web 検索は何らかの情報を得るための有効な手段となっているが、知りたい情報を探し出すことはそれほど容易ではない。

自然言語による質問文に対して、情報源として Web を用い、回答を Web 文書から抽出して提示する質問応答システムが研究されている。質問応答システムは、既存の Web 検索と異なり、自然な言葉で質問を入力する点と、システムが自動的に回答を抽出して回答する点で優れており、ユーザの負担を軽減することが期待できる。

How-to 型質問を対象としたシステムは、What, When, Who, Where 型の質問を対象としたシステムと異なり、文で回答する必要がある。そのため、回答範囲の特定が難しい。また、質問応答システムでは、内容の類似している回答を抽出することが多いため、どの回答が適しているかをユーザに判断してもらわなければならない。特に文で回答するシステムでは、名詞で回答するシステムに比べて回答が長いので、負担が大きくなる。

本研究では、Web を知識源に用い、何らかの状況に置かれたユーザに対してどうすれば良いかを回答する質問応答システムについて検討する。

2. How-to 型質問応答システム

2.1 行動表現

筆者らは、方法説明を含む Web 文書の多くが箇条書き表現を用いていることを利用して、How-to 型質問に対する回答範囲の特定に箇条書きを表現が有効であることを示した [1]。しかし、類似した箇条書き表現が多く抽出されてしまい、正しい回答を選択することが困難になった。正しい回答を選択するためには、回答候補の passage に現れる名詞や動詞に着目する必要がある。特に How-to 型質問応答システムでは回答として行動を返す必要があるため、行動に注目する必要がある。そこで、三原ら [1] が着目した「名詞+助詞+動詞」という形式をとる行動表現を回答候補の passage 選択に利用する。行動表現は CaboCha を用いて係り受け解析を行い、「名詞+助詞+動詞」という係り受け構造を抽出する。

2.2 システムの構成

本システムでは、「しゃっくりが止まらない」などユーザが置かれている状況を質問文としてを用いる。なお、How-to 型の質問に焦点を絞っているため、一般的な質問応答システムのように、入力された質問文から質問タイプの同定は行わない。質問文から名詞句と動詞を抽出してクエリとし Web 検索を行い、取得した文書から箇条書きとそれに続く文を passage として抽出する。抽出

した passage から行動表現を抽出し、各ページごとに行動表現の流れを取得する。各ページごとの行動表現の流れを比較することで、回答を抽出する。

本システムの構成を図 1 に示す。

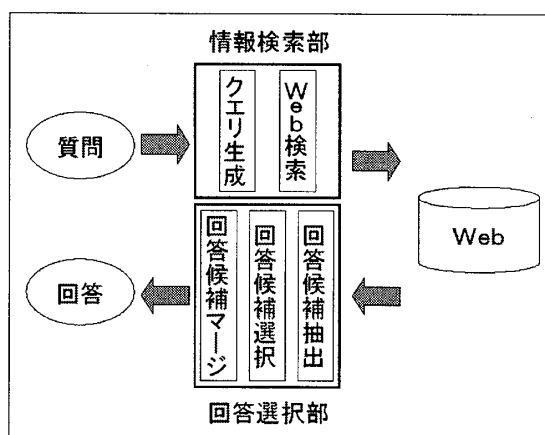


図 1: システムの構成

2.2.1 情報検索部

質問文から助詞を削除して検索語を生成し、検索エンジンを用いて文書の取得を行う。検索エンジンは Google を用いた。なお、How-to 型質問応答システムに限定し、質問タイプの解析は行わない。

2.2.2 回答選択部

- 回答候補抽出
各文章から箇条書きとそれに続く文を passage として抽出する。箇条書きかどうかの判別は、箇条書きを表すタグと文頭の数字や記号を手がかりにする。
- 回答候補選択
passage は箇条書き表現を抽出しているだけなので、その内容が回答として適切かどうか分からない。そこで、行動表現を手がかりにして回答候補となる passage を抽出する。

– キーワードを含む passage の抽出

質問文から行動表現を抽出し、類似した行動表現が passage に存在するかどうか調べる。行動表現が、完全に一致するか、名詞のみが一致するか、動詞のみが一致する場合にその passage と、その前後の passage に距離に応じてスコアを与える。

Answer Selection for How-to Type Question Answering System.

Daisuke Sugiyama, Ikuo Tahara.

*Graduate School of Tokyo University of Science.

†Tokyo University of Science.

- 行動表現の頻度
頻度の高い語は重要な語であるという考えから、全ての passage から行動表現を抽出してその出現回数が多い行動表現を含んだ passage を回答候補として高いスコアを与える。なお、質問文から抽出した行動表現と、動詞は除く。
- passage 内に現れる推奨表現
推奨表現は有効な対処法を述べる時、禁止表現は行ってはならない行動を述べる時に用いられることが多いことから、「～しましょう」のような推奨表現と「～してはいけない」のような禁止表現を伴う passage にスコアを与える。

表 1: 推奨表現の様相

様相	文末表現の一例
依頼	～ください
勧誘	～ましょう
勧め・忠告	～たほうがいい
命令	～なさい
義務・必要・不必要	～てはいけない
希望	～てほしい

スコアは、上記の基準を用い以下の式により求める。

$$s(x) = \frac{1}{a_x} \cdot w + \frac{b_x}{b} + c_x \quad w = \{1.0, 0.5, 0.3\} \quad (1)$$

ここで、 x は passage に含まれる行動表現である。 a_x はキーワードを含む passage との距離を表し、 w は質問文中の行動表現と完全一致、動詞、名詞が一致する場合にそれぞれ 1.0, 0.5, 0.3 を与える。 b_x は行動表現の頻度、 b は抽出した行動表現の総数を表す。 c_x は推奨表現の有無により 1 か 0 を与える。

passage P のスコアは、以下の式で求める。

$$s(P) = \sum_{x \in P} s(x) \quad (2)$$

● 回答候補マージ

回答候補として各文書から抽出した passage は同一の内容である可能性が高い。そこで、各ページの passage から行動表現の流れを抽出し、同一の内容である場合それらをひとつにまとめる。

1. 行動表現にスコアを与える

各ページからスコアの高い passage を抽出し、行動表現を抽出する。抽出する行動表現は回答として適切な行動表現であり、その判定には、passage のスコア、行動表現の出現頻度、推奨表現の有無でスコアを取る。なお、質問文から抽出した行動表現と動詞は抽出対象から外す。

$$s(x) = P + \frac{b_x}{b} \cdot w + c_x \quad w = \{1.0, 0.5, 0.3\} \quad (3)$$

2. 行動表現の流れ抽出

各ページで上部から passage のペアを作り、passage 間で行動表現のペアを作り、スコアの高い行動表現のペアを抽出する。

3. 回答を抽出

各ページから抽出した行動表現の流れを比較し、スコアの高い行動表現の流れを含む passage を回答として抽出する。

3. 評価実験

ユーザの置かれた状況を入力し、それに対する回答として、どの行動表現の流れが有効であるか正解率を求める。取得 URL は上位 30 件とした。質問文は「蜂に刺された」、「しゃっくりが止まらない」など 30 件使用した。この実験の正解率を表 2 に示す。

表 2: 正解率

回答順位	件数 (件)	正解率 (%)
1 件目	12	40.0
3 件目	22	73.3
5 件目	23	76.7
10 件目	24	80.0
30 件目	24	80.0

4. おわりに

本手法は、How-to 型質問に対する回答として簡条書きを用い、行動表現を用いることで正解となる一連の行動を抽出することが出来る。今後、被験者を用いた実用性を含めた評価がさらに必要である。

参考文献

- [1] 杉山大輔, 延澤志保, 太原育夫: “簡条書き表現に着目した質問応答システム,” FIT (情報科学技術フォーラム), E-004, pp.137-138, 2007.
- [2] 三原英理, 藤井敦, 石川徹也: “World Wide Web を用いたヘルプデスク指向の質疑応答システム,” FIT (情報科学技術フォーラム), E-021, pp.163-166, 2005.
- [3] 日高直哉, 榎井文人: “質問応答における回答絞り込み手法の比較,” 第 17 回人工知能学会全国大会発表論文集, 3C4-8, 2003.
- [4] 佐藤充, 石下円香, 森辰則: “Web 文書を情報源とする記述的な回答が可能な質問応答システム,” 言語処理学会第 14 回年次大会発表論文集, A5-3, pp. 1009-1012, 2008.