E-004

箇条書き表現に着目した質問応答システム
Question Answering System by Exploiting the Description in the Form of Itemized Expressions

杉山 大輔 * 延澤 志保 † 太原 育夫 ‡
Daisuke Sugiyama Shiho Nobesawa Ikuo Tahara

1. まえがき

現在，インターネットの普及により，様々な情報が Web 上に存在している．そのため，Web 検索は知りたい情報を得るための有効な手段になっている．Web から情報を取得する手段として検索エンジンを用いることが一般的であるが，既存の Web 検索は調べたい内容を的確にキーワードで表現する必要がある．また，検索結果はキーワードを含む文書を提示するだけなので，必要とする情報を入手するためには，ユーザ自身が検索結果の文書から回答を見つけ出さなければならない．

自然言語による質問文に対して，情報源として Web を用い，Web 文書から回答を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，検索エンジンを用いた情報検索に比べ，ユーザが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探し出して答える点で優れており，ユーザの負担を軽減することが期待できる．

質問応答システムの多くは What, When, Who, Where 型の WH 質問文に対して名詞及び名詞句を回答するもので，事実を回答するものである．しかし，How-to 型などの質問文に対しては，回答として単語は適切ではなく，文章を回答する必要がある．このような文章での回答を求める質問に対する質問応答システムの研究はあまり進んでおらず，How-to 型の質問応答システムとしては「名詞+助詞+動詞」という行動表現を抽出して回答するシステムがある [1]．

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 方法説明を含む文書の特徴

麻野間らは，方法説明を含む文書がどのような特徴を持っているかを調査した [2][3]．すなわち，「梅干　漬け方」のような方法を尋ねる質問をクエリとして与え，方法説明が含まれる文書構造を調査し，その結果，方法説明を含む文書の多くが箇条書き表現を取っていることを明らかにした．

How-to 型質問を対象としたシステムは，名詞や名詞句を回答として返す WH 型質問応答システムと異なり，回答部分の特定が難しい．そこで，麻野間らの調査した結果を元に How-to 型質問応答システムの回答として Web から抽出した箇条書きを用いる手法を提案する．Web から箇条書きを抽出する先行研究として，抽出した<li>タグの付与されたテキストが手順を表すかどうかを判別したもの [4] や，料理レシピから箇条書きを抽出して利用するもの [5] がある．

2.2 システムの構成

本システムでは，質問文として「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表す文を用いる．なお，How-to 型の質問に焦点を絞っているので，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分類は行わない．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抽出する．このとき，ノイズとして含まれる箇条書きが回答の上位に現れないように，箇条書きの文末表現に注目し，方法説明性を有している箇条書きを多く含む Web 文書の箇条書き群を回答として返す．

本システムの構成を図 1 に示す．

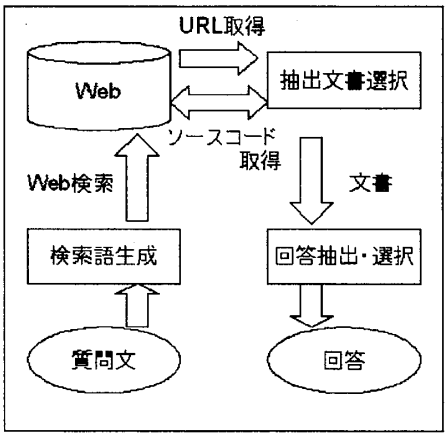

図 1: システムの構成

2.2.1 検索語生成

ユーザは現在置かれている状況や悩みを質問文として入力する．システムはこの入力された質問文を形態素解析し，助詞を取り除き，残った名詞句と動詞を検索語とする．例えば，ユーザが「しゃっくりが止まらない」と入力した場合，検索語は「しゃっくり　止まらない」となる．

2.2.2 Web 検索

2.2.1 節で得られた検索語を用いて Web 検索を行う．検索エンジンは Google を用いる．検索結果のページか

*東京理科大学大学院理工学研究科情報科学専攻
†武蔵工業大学知識工学部情報科学科
‡東京理科大学理工学部情報科学科

らソースコードを取得し，URL を取得し，Web ページのソースコードを取得する．

2.2.3 抽出文書選択

2.2.2 節で得られた Web ページに方法説明を含むような語が含まれているかチェックし，含まれていなければ箇条書きの抽出対象から外す．方法説明を表す語として「方法」や「対処」などを用い，文末表現として「～ましょう」,「～なくてはいけない」など，ユーザに行動を促す表現を用いた．文末表現の一例とその様相を表 1 に示す．

表 1: 方法説明の文末表現の様相

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

2.2.4 回答抽出

Web ページのソースコードから passage を抽出する．箇条書きに続く文は，直前の箇条書きの内容を詳しく書いたものであることが多いので，passage の範囲は，箇条書きから次の箇条書きまでの文とする．

- 箇条書き部分の抽出
  - 箇条書きを表すタグの場合
     タグで囲まれている文．
  - 箇条書きを表すタグでない場合
    文頭の数字や記号を手がかりにする．判別には形態素解析を行い，文頭が記号，または数字列＋空白や記号，未知語の時，箇条書きと判定する．
- 箇条書きに続く文の抽出
  次に箇条書きと判定されるまで，箇条書きに続く文とする．

2.2.5 回答選択

抽出した passage に対してスコアをつけて昇順に並べる．スコア付けには 2.2.3 節で用いた方法説明に現れやすい語や，ユーザに行動を促す文末表現を用いる．

- passage のスコア付け
  方法説明を表す表現が現れたら加点し，合計点を P とおく．passage の文字数を N とおき，passage のスコアを次の式で求める．

$$score(n) = \frac{P}{N} \quad (1)$$

ここで、合計点を文字数で割るのは，文字数が少なく方法説明を現す語が多用されているもの，つまり方法を簡潔に現しているものを回答の上位にするためである．

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，どの程度箇条書き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「交通事故を起こした」,「しゃっくりが止まらない」など 30 件使用した．方法説明性が高い箇条書きを多く含む Web 文書の箇条書きすべてを回答として出力しているため，方法説明を表さない箇条書きを含むもののうち，1 件目に回答が現れたのが 30 %，上位 10 件以内に回答が含まれるのが 80 % であった．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	9	30.0
3 件目	20	66.6
5 件目	21	70.0
10 件目	24	80.0
30 件目	24	80.0

4. おわりに

本研究では，How-to 型質問に対する回答として箇条書きを用いることが有効であることを示した．回答選択部分では方法説明のみに注目しているため，今後は，質問文との関連性などを考慮していく予定である．

参考文献

[1] 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.

[2] 麻野間直樹，古瀬蔵，片岡良治: “How-to 型質問応答の実現に向けた質問回答文書の特徴分析, ”情報処理学会研究報告，2006-NL-168，pp.55-60，2006.

[3] 麻野間直樹，古瀬蔵，片岡良治: “文書構造と言語表現の分析に基づく方法説明抽出, ”言語処理学会第 12 回年次大会発表論文集，pp.324-327，2006.

[4] 武智峰樹，徳永健伸，松本裕治，田中穂積: “WWW ページからの手順に関する箇条書きの抽出, ”情報処理学会論文誌:データベース，Vol.44，No.SIG12，pp.51-63，2003.

[5] 田島幸恵，奥村学: “Web 上の料理レシピの抽出とその利用, ”言語処理学会第 11 回年次大会発表論文集，pp.65-68，2005.