HTML文書を対象とした質問応答システムにおける回答抽出方法

3V−2

大沼宏行　池野篤司
沖電気工業株式会社　研究開発本部　情報サービスシステムラボラトリ

1. はじめに

近年、膨大な情報の中から、ユーザが求める情報をピンポイントで回答する技術として、質問応答システムが研究されている。例えば、佐々木ら[1]は、新聞記事を対象として質問応答システムを作成し評価している。これらのシステムでは形態素解析、係り受け解析など様々な処理を利用して、回答を抽出している。

一方、インターネット上の情報から回答を抽出する場合には、様々なレイアウトの文書がヒットする。したがって、それらを解析し、適切に回答を選別することが重要になる。

2. 方針

賀沢ら[2]のシステムでは、回答の抽出方法として、質問文中のキーワードを全て含む範囲を計算し、その範囲内で回答を探す方法が提案されているが、本稿で対象とするＨＴＭＬ文書などの構造化された文書では必ずしもうまくいくわけではない。なぜなら、ＨＴＭＬ文書では、見出し語、入れ子構造、表構造などがあり、すべてのキーワードが近傍にあるとは限らないからである。例えば、「東京で行われる○○のセミナーの会場はどこですか？」という質問があった場合には、図１の文書のように、回答候補（図１の下線部）の近傍にキーワード(図１の太字部)が存在しない場合がある。つまり、キーワードが、入れ子構造の外側にあったり、表の先頭行にあったり、見出し行になって いる場合がある。

そこで、我々は、ＨＴＭＬ文書のタグ情報を利用して、キーワードと回答候補の位置関係を考慮し、回答を選択する。

Answer Extraction of Question and Answering System on HTML Documents
Hiroyuki OHNUMA and Atsushi IKENO
Service Media Laboratory, Corporate Research and Development Center, Oki Electric Industry Co.,Ltd.

3. システム構成

図２に、本稿で述べるシステムの構成を示す。構成要素として、ユーザインターフェース、質問解析部、文書検索部、回答選択部がある。質問解析部は、賀沢らの手法を用い、質問文中の名詞を抽出しキーワードにしたり、回答の固有表現パターン（日付、住所、組織名のうち、回答をどれにすればいいか）を決定する。例えば、「東京で行われる○○のセミナーの会場はどこですか？」という質問では、キーワードを名詞とすると、「東京」「○○」「セミナー」「会場」がキーワードとなる。また、回答の固有表現パターンは、疑問詞「どこ」に対して住所パターン（△△市○○区××）を抽出することにする。文書検索部は、質問解析部で抽出されたキーワードで、インターネット上の文書を検索する。回答選択部は、文書検索部で検索された個々の文書に対して、キーワードと回答候補の位置関係を計算し、回答候補をスコアリンクし、回答を選択する。特に回答選択部の処理については、４章で詳細に述べる。

```
1  <HTML><HEAD><TITLE>○○セミナー</TITLE></HEAD>
2  <BODY>
3  ...
4  <TABLE BORDER=1>
5  <TR><TD COLSPAN=3><B>東京</B></TD></TR>
6  <TR><TD rowSpan=2 width=65>開催日</TD>
7  <TD><B>平成13年3月17日</B><BR>【終了しました】</TD>
8  <TD>10:00~12:30</TD></TR>
9  <TR><TD><B>平成13年4月1日</B><BR>【終了しました】</TD>
10 <TD>10:00~12:30</TD></TR>
11  ...
12 <TR><TD width=65>会場</TD>
13 <TD COLSPAN=2>中央区○○1-1-1</TD></TR>
14 </TABLE>
15 ...
16 </BODY></HTML>
```

図１：文書例

4. 回答抽出方法

4.1 基本範囲

賀沢らの手法では、回答を調べる最小範囲として、文を単位にしている。それと同様に、本手法で取り扱うＨＴＭＬ文書でも回答を調べる最小範囲を決めておく。この最小範囲を基本範囲と呼ぶ。

基本範囲を次のように定義する。

(1)「。」で区切られる一つの文

(2)次の条件で決められる範囲

ただし、ブロックは、<BODY></BODY>内にあり、<P><HR><TABLE><UL><OL></TABLE></UL><LI></OL>と<Hn></Hn>(1≦n)で区切られる範囲とする。

(3)タイトル内<TITLE>、表の同じ行<TR>

結果として、<TD><CENTER>
などのタグや<FONT><B>などのタグは無視する。

4.2 キーワードと回答候補の位置関係

回答の選択条件として有効なキーワードと回答候補の位置関係を次の場合に分ける。

[(範囲 1) キーワードと回答候補がかなり近く、強い影響関係にある場合]

(1) 基本範囲内にキーワードと回答候補がある。例えば、図１の 12 行目「会場」と 13 行目の回答候補の関係である。

[(範囲 2)キーワードと回答候補がやや近く、比較的強い影響関係にある場合]

(2) キーワードと回答候補が同じ箇条書き内にある。

(3) キーワードが表の 1 行目にあり、回答候補と同じ表内にある。例えば、図１の 5 行目「東京」と１３行目の回答候補の関係である。

(4) キーワードと回答候補が同じ表内にある。((3)以外の場合)

(5) 箇条書きの入れ子構造になっており、入れ子の外にキーワードがあり、入れ子の中に回答候補がある場合。

[(範囲 3)キーワードと回答候補が、見出しとその内容という関係にある場合]

(6) キーワードが章節名にあり、回答候補が、その章または節内にある。

(7) キーワードがタイトルにあり、回答候補が、その本文内にある。例えば、図１の１行目「○○セミナー」と１３行目の回答候補の関係である。

4.3 回答抽出方法

回答抽出処理の処理手順は、次の通りである。

[Step.1] 検索された文書に対して、キーワードと、日付や住所などの固有表現にタグを付加する。

[Step.2] キーワードと回答候補の位置関係を計算しやすくするために、Step.1 の文書のタグの包含関係を解析し、影響関係を表す木構造を作成する。

[Step.3] キーワードと回答候補の間の関係が、4.2 節の(1)～(7)のうち、いずれであるかを計算する。

[Step.4] (1)～(7)の関係に応じて、各回答候補にスコアをつける。

[Step.5] スコアがつけられた回答候補を比較し、検索されたすべての文書のうち、最も高いスコアの回答候補を回答にする。

5. おわりに

本稿では、HTML 文書を対象とした質問応答システムにおいて、キーワードと回答候補の位置関係を HTML タグを利用して計算する回答抽出方法を述べた。今後、本抽出方法の評価に取り組む。

参考文献

[1] 佐々木裕ら: 質問応答システムの比較と評価, 信学技報, NLC2000-24, pp.17-24 (2000).

[2] 賀沢秀人,加藤恒昭: 意味制約を用いた日本語質問応答システム,情報処理学会研究報告, 2000-NL-140, pp.173-180 (2000).

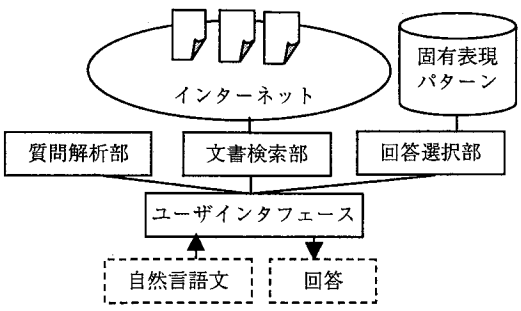

図２：システム構成