ウェブ文書を知識源とした曖昧な質問に対する質問応答

長内　亘　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{s0610019, kshirai}@jaist.ac.jp

1 はじめに

我々は曖昧な質問を取り扱う質問応答システムに関する研究に取り組んでいる [3, 5]．曖昧な質問とは，ユーザの質問文中のキーワードの意味が曖昧であるために解答を一つに絞ることができない質問を指す．本研究では，曖昧な質問に対して複数の解答をリストとして提示する．また，解答とともに，質問文中の曖昧なキーワードの意味を限定する表現 (以下，これを「限定表現」と呼ぶ) も提示する．例えば,「ワールドカップの優勝国はどこですか」という質問は，ワールドカップにはサッカー，ラグビーなどの種類があるという意味で曖昧であり，競技の違いに応じて解答が複数存在する．このとき，本研究では以下のような解答リストをユーザに提示する．

・イギリス (ラグビーのワールドカップ)
・ブラジル (サッカーのワールドカップ)
・ノルウェー (スキーのワールドカップ)

ここで,「ラグビー」「サッカー」「スキー」はワールドカップの意味を限定する限定表現である．

複数の解答を出力するリスト型質問応答システムに関する研究は既に数多く行われている [1, 2]．本研究では,「四大文明とは何ですか」のような単に複数の解答を尋ねる質問ではなく，上記のような曖昧な質問に対して適切な解答リストを提示することを目的とする．

我々のこれまでの研究 [3, 5] では，質問に対する解答を探し出す知識源として新聞記事を用いた．これに対し，本論文では知識源としてウェブを用いる [4]．新聞記事と比べて，ウェブはテキストの量がはるかに多く，またより多様な情報が存在するなど，質問応答システムの知識源として望ましい点も多い．また，曖昧な質問に対して解答リストを提示する我々のこれまでの手法は，知識源となるテキスト中の文を解析し，解答リストを動的に生成していた．これに加え，本研究ではウェブページにおける表に着目する．質問によっては，ウェブにおける様々な表の中に，その質問に対する解答リストとしてふさわしいものが存在するときがある．本論文では，ユーザに提示する解答リストとなりうる表を発見する手法を提案し，従来のテキスト解析に基づく手法と併用する．

2 システム概要

提案システムにおける処理の流れを図 1 に示す．

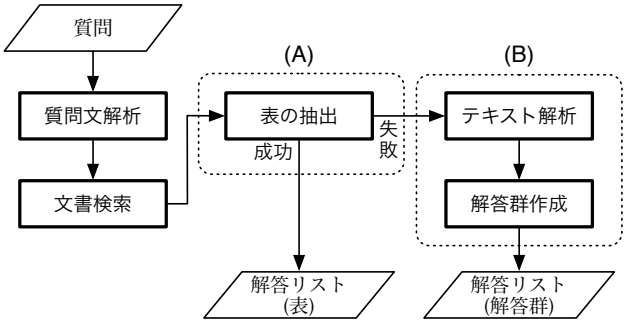

図 1: 提案システムの概要

まず，ユーザから入力された質問文の解析を行い，質問タイプの同定とキーワードの抽出を行う．ただし，キーワードはプライマリキーワードとセカンダリキーワードの 2 つのタイプに分ける．プライマリキーワードは解答と最も関係の深いキーワード 1 つであり，例えば質問文中の主題にあたる名詞などが該当する．質問文中に含まれるそれ以外のキーワードは全てセカンダリキーワードとする.「ワールドカップの優勝国はどこですか」という質問の例では，プライマリキーワードは「優勝国」，セカンダリキーワードは「ワールドカップ」となる．次に，キーワードをクエリとしてウェブページを検索する．検索エンジンは Tsubaki [6] を利用し，検索結果の上位 100 件のウェブページを獲得する．

獲得されたウェブページの集合から，ユーザに提示する解答リストを 2 つの手続きによって作成する．まず，検索されたウェブページの中に解答リストとしてふさわしい表があれば，それを抽出して出力する (図 1 の (A))．解答リストとするべき表がみつからなかった場合には，ウェブページ内のテキストを解析して複数の解答候補を抽出し，それらをまとめて解答群を作成し，ユーザに提示する (図 1 の (B))．以下，図 1(A) の処理の詳細については 3 節で，図 1(B) の処理の詳細については 4 節で述べる．

3 解答リストを含む表の抽出

本節では，ユーザの曖昧な質問に対する複数の解答を含む表をウェブページから抽出する手法について述べる．以下にその手続きを示す．

1. table タグで定義されている表を検出する．ウェブページにはテキストや画像などで表が記述されている場合もあるが，これらは抽出の対象外とする．
2. 表の 1 行目または 1 列目にあるセルとユーザの質問文におけるプライマリキーワード k_p が一致するかをチェックする．一般に，表の 1 行目または 1 列目は何らかの属性を表わすとみなせる．もし，k_p と表の属性が一致していれば，その表は質問の解答を含む可能性が高い．
   また，表のセル内の文字列と k_p が完全に一致しなくても，以下の条件を満たすときには，その文字列はプライマリキーワード k_p を表わすとみなした．
   - セル内の文字列が複合名詞である (動詞などを含まない)．
   - セル内の文字列の末尾，または最後の 1 文字を除く末尾が k_p と一致する．
3. 全てのセカンダリキーワードが (1) 表のキャプション，(2) 表の前にある 3 つのセグメント [1]，(3) そのページの title タグの中 [2]，のいずれかの場所に存在するかをチェックし，存在しないときはその表を除外する．
4. 2. で検出したセルが表の 1 列目にあるとき，そのセルと同じ行にあるセル内のテキストが解答を含むかをチェックする (図 2(a))．同様に，2. で検出したセルが表の 1 行目にあるとき，そのセルと同じ列にあるセル内のテキストが解答を含むかをチェックする (図 2(b))．セルが 1 行 1 列目のときは両方の可能性をチェックする．
   表の一行または一列が解答を含むかどうかのチェックは以下のように行う．まず，セル内のテキストを南瓜 [3]で解析し，固有表現タグを割り当てる．その固有表現タグがユーザの質問の解答タイプと一致しているかをチェックする．一行または一列内のセルのうち，固有表現タグと解答タイプが一致している割合を調べ，それが 0.3 以上のときにはその表を抽出し，ユーザに提示する．

[1] セグメントの定義については 4 節の冒頭で述べる．
[2] title タグにあるキーワードはページ全体のトピックを表わすと考えられる．ここでは，キーワードが title タグにある場合は，そのページは質問と関連が深いとみなし，ページ内の表を抽出の対象とした．
[3] http://chasen.org/%7Etaku/software/cabocha/

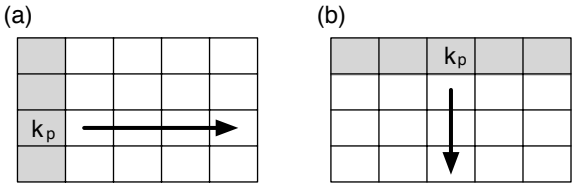

図 2: 複数の解答を含む表の検出

上記の手続きで抽出された表には，解答だけでなく，質問文中の曖昧なキーワードに対する限定表現も含まれていると期待できる．ユーザは，自分の質問が曖昧であることを意識していなくても，表を見ることで自分の質問の曖昧性に気づく可能性もある．

ひとつの質問に対して複数の表が得られたときはその全てを出力する．複数の表を何らかの基準でランク付けして，最も適切な表をひとつ選択することが望ましいが，本研究では行っていない．今後の課題としたい．

4 テキスト解析による解答群の作成

本節では，テキスト解析に基づいて，ユーザに提示するべき解答リストを動的に生成する手法について述べる．なお，ここで述べる手法は文献 [5] で提案した手法に準じ，これをウェブを知識源とする場合に対応するように改変したものである．以下はその概要を述べる．詳細は文献 [4] を参照していただきたい．

まず，キーワード検索によって得られたウェブページをおおまかなセグメントに分割する．具体的には，以下に挙げる HTML タグによってウェブページを分割し，分割された個々のテキスト領域をセグメントとする．

address, blockquote, div, dl, h1, h2, h3, h4, h5, h6, hr, ol, p, pre, table, ul, noframes, noscript, dir, menu

次に，以下のいずれかの条件を満たすセグメントを抽出する．

- 全てのキーワードが含まれるセグメント
- セカンダリキーワードのいずれかがそのセグメントのウェブページ全体の title タグに含まれ，かつプライマリキーワードを含む残りのキーワードが含まれるセグメント

得られたセグメントを南瓜で解析し，固有表現タグや文字種の情報などを手がかりに解答候補を抽出する．

次に，複数の解答候補 a_i のそれぞれに対して，それを含む元のセグメントから (a_i, k_j, s_k) という 3 つ組を抽出する．k_j はユーザが入力した質問文内のキーワードであり，s_k はキーワード k_j の意味を限定する限定表現である．限定表現はあらかじめ用意されたいくつかの抽出

パタンによって抽出する．例えば，キーワードの直前・直後にある名詞や，キーワードと意味的関連度の高い名詞が限定表現として抽出される．次に，(a_i, k_j, s_k) の集合から，k_j が共通でかつ限定表現 s_k に何らかの共通の属性 $(attr)$ が存在する部分集合を発見する．共通属性の例としては，末尾 N 文字が共通 (N=1,2,3)，意味クラスが共通,「数+助数詞」という表現であることが共通，などがある．このように発見された部分集合を

$$AG(k, attr) = \{ (a_i, s_i) \} \quad (1)$$

という形式で表現し，解答群とする．解答群 AG において，k は 3 つ組の部分集合に共通するキーワードであり，これはユーザの質問文に含まれる曖昧なキーワードに該当する．一方，s_i はキーワード k の限定表現，$attr$ は s_i が共通して持つ属性である．

解答群 AG は一般に複数得られるため，(1)AG における限定表現や解答候補の異なり数，(2) 限定表現の共通属性 $attr$ のタイプ，(3) 解答候補の信頼度，(4) キーワードと限定表現の関連度，などに応じてスコアを付ける．最大のスコアを持つ解答群をユーザに提示する解答リストとする．

5 予備実験

本節では提案手法を評価する予備実験について報告する．まず，テストデータとして曖昧な質問を 30 問用意した．これらの質問は著者によって作成した．

5.1 表の評価

3 節で述べた手法を用いて表を抽出し，その表が質問に対する解答リストとして適切であるかを人手で判定した．結果を表 1 に示す．

表 1: ウェブページから抽出された表の評価

抽出された表の数	24
精度	88%
再現率	46%

表 1 における精度は，システムによって出力された表のうち正解とみなせる表の割合である．一方，再現率は以下のように求めた．今回の実験では，1 つの質問に対して検索結果の上位 100 件のウェブページを用いたため，合計 3,000 個のウェブページが表の抽出処理の対象となる．これらのウェブページを人手でチェックし，解答リストとしてふさわしい表を抽出した．再現率は，このようにして得られた表のうち，実際にシステムによって取り出された表の割合である．

30 個の質問に対して，表が 1 つ以上得られた質問の数は 10 であり，また再現率が低いことから，提案手法は本来取り出すべき多くの表の抽出に失敗している．その主な要因は以下の通りである．

- キーワードと表の属性の表記が一致しない．(キーワードが「受賞者」で表の属性が「氏名」の場合など)
- 属性がない表，あるいは属性が必ずしも 1 行目，1 列目にない複雑な表に対して，キーワードと表の属性とのマッチングに失敗する．
- 固有表現解析の失敗により，解答が並んだ行や列の認識に失敗する．

一方，表の抽出の精度は高く，誤った表を抽出した事例は 3 件であった．質問応答システムでは，正解となる全ての表を抽出する必要はなく，正しい表をひとつ見つけてユーザに提示すれば十分である．したがって，再現率よりは精度が重視される．表 1 に示した実験結果は，上記のような観点からは望ましいといえる．

実際に解答リストとして抽出された表の例を図 3 に示す．この表は,「全英オープンで優勝したのは誰ですか?」という質問に対して得られた．

全英オープンゴルフ・歴代優勝チャンピオン

年	優勝者	出身国	開催地
2006	タイガー・ウッズ	米	ロイヤルリバプール・ゴルフクラブ
2005	タイガー・ウッズ	米	セント・アンドリュース
2004	トッド・ハミルトン	米	ロイヤルトルーン
2003	ベン・カーティス	米	ロイヤルセントジョージズ
2002	アーニー・エルス	南ア	ミュアフィールド

図 3: 抽出された表の例

「全英オープン」には開催年の曖昧性があり，それに対する複数の解答が得られていることがわかる．この場合，曖昧なキーワードは「全英オープン」であり，その限定表現は「年」である．本来なら表中の「年」が限定表現であることを特定し，すなわち「全英オープン」には開催年という観点で曖昧性があることを特定した上でユーザに提示すべきであるが，本研究では行っていない．また，提案手法によって抽出された正解の表を調べたところ，図 3 のような大会の開催年や開催回数に関する曖昧性を表わす表がほとんどであった．全英オープンの例では，ゴルフやテニスといった競技に関しても曖昧性があるが，そのような観点でまとめた表は今回の実験では抽出できなかった．開催年以外の曖昧性を表わす表とし

て,「アカデミー賞を受賞したのは誰ですか?」という質問に対し，賞の部門毎に複数の解答を含む表が得られたが，その表は「日本アカデミー賞」の受賞者の一覧を示しており，今回の実験では不正解とした．

5.2 解答群の評価

4 節で述べた手法を用いて解答群を作成した．ここでは解答群の作成手法を評価するため，システムが抽出した解答候補のうち正解となるものを人手で選別し，解答群を作成した．スコアの上位 10 件の解答群を人手でチェックし，それが質問に対する解答リストとして適切であるかを判定した．結果を表 2 に示す．

表 2: 生成された解答群の評価

スコア 1 位	13(43%)
10 位以内	22(73%)
正解解答群の平均順位	2.1

表 2 において,「スコア 1 位」はスコア最大の解答群が解答リストとして適切とみなせる質問の数とその割合,「上位 10 位」はスコアの上位 10 位の解答群の中に正解となる解答群が含まれる質問の数とその割合,「平均順位」は正解の解答群の順位の平均である．全体の 4 割程度の質問に対して，解答リストとしてふさわしい解答群を抽出できた．

「シドニー五輪の柔道の金メダリストは誰ですか?」という質問に対して生成された解答群の例を以下に示す．

$AG($ 金メダリスト, 数+キロ級 $) =$
$\{$ (井上康生, 男子１００キロ級),
(井上康生, 五輪１００キロ級),
(田村亮子, 女子４８キロ級) $\}$

上記解答群中の「数+キロ級」は，限定表現が数量表現の後に「キロ級」が続くという共通の属性を持つことを表わす．解答に重複はあるが,「金メダリスト」というキーワードに階級の曖昧性があることが示されている．ただし，生成された解答群を調べたところ，5.1 項で抽出された表と同様に，大会の開催年や開催回数の曖昧性を表わすものが多かった．

5.3 組み合わせ手法の評価

3 節と 4 節の手法を組み合わせた提案手法によって解答リストを作成した．評価結果を表 3 に示す．

表 3 において,「表」は解答リストとしてウェブページ内の表を提示する手法 (3 節),「解答群」はテキスト解析に基づいて生成された解答群を提示する手法 (4 節),「併用」はまず表を抽出し，それに失敗したときに解答群を生成するという方式で両者を併用する提案手法 (図 1) を表わす．一方，表 3 (A) は何からの解答リストが得られた質問の数，(B) はシステムによって得られた解答リスト (表ならば抽出した表全て，解答群ならスコアの上位 10 位の解答群) の中に正解を含む質問の数，(C) はスコアの上位 1 位の解答リストが正解となる質問の数を表わす[4]．表 3 の結果から，(B),(C) の場合ともに，2 つの手法を組み合わせることで正解が得られる質問の数が増えたことがわかった．

表 3: 解答リストの手法別の評価

	表	解答群	併用
(A) 出力	10	30	30
(B) 正解を含む	9	22	25
(C) 正解が 1 位	9	13	17

6 おわりに

本論文では，ウェブを知識源とし，曖昧な質問が入力されたときに複数の解答からなるリストをユーザに提示する質問応答システムについて述べた．現状では，得られる解答リストの多くは大会の開催年や開催回数の曖昧性を反映したものであり，ユーザの多様な質問に対応できているとは言い難い．より多様な曖昧性を検出し，適切な解答リストを作成する方法を検討することが今後の課題である．

参考文献

[1] 石下円香, 森辰則. 優先順位型質問応答の解スコア分布に基づくリスト型質問応答. 情報処理学会自然言語処理研究会, Vol. 2005, No. 94, pp. 41–47, 2005.

[2] 加藤恒昭, 桝井文人, 福本淳一, 神門典子. リスト型質問応答の特徴付けと評価指標. 情報処理学会自然言語処理研究会, Vol. 2004, No. 93, pp. 115–122, 2004.

[3] 松本匡史, 白井清昭. 質問の曖昧性を検出し複数の解答を提示する質問応答システム. 言語処理学会第 12 回年次大会, pp. 935–938, 2006.

[4] 長内亘. ウェブを知識源としたユーザの曖昧な質問に対する質問応答. Master's thesis, 北陸先端科学技術大学院大学, 3 2008.

[5] 坂本篤史, 白井清昭. 対話型質問応答システムにおける曖昧な質問に対する問い返し文の生成. 言語処理学会第 13 回年次大会, pp. 1006–1009, 2007.

[6] 新里圭司, 柴田知秀, 河原大輔, 黒橋禎夫. 大規模日本語ウェブ文書を対象とした開放型検索エンジン基盤の構築. 言語処理学会第 13 回年次大会, pp. 1117–1120, 2007.

[4] 表については順位付けを行っていないので，抽出された表の中に正解が含まれる質問の数を示した．