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

長内 亘 (0610019)

北陸先端科学技術大学院大学 情報科学研究科

2008 年 2 月 7 日

キーワード: 質問応答, 質問の曖昧性, ウェブ, ウェブページの表.

本論文はウェブ文書を知識源としたオープンドメインなリスト型質問応答システムについて述べる．このシステムは，ユーザの質問が曖昧であるときに，その曖昧性を検出し，複数の解答をリストとして提示する．本研究における「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 1 つに絞ることができない質問を指す．例えば,「ワールドカップで優勝した国はどこですか」という質問は，ワールドカップにはサッカーやスキー，ラグビーなど様々なスポーツの種類が存在し，その種類によって解答が異なるという意味で曖昧である．本システムはこのような質問に対して,「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)」,「イギリス (ラグビーのワールドカップ)」のような曖昧なキーワードの意味とそれに対応する解答のリストを提示する．先行研究が解答を得るための知識源として新聞記事を用いていたのに対し，本論文では知識源としてウェブを用いる．また，上記のような解答リストを抽出する手法として，ウェブページにおける表に着目する．ウェブページの表の中にはその質問に対する解答リストが存在する場合があるからである．本研究では，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方法を提案する．

曖昧な質問に対して解答リストを提示するシステムの処理の流れは以下の通りである．まず，ユーザの質問文を解析して，キーワード，解答タイプ，キーワードタイプを抽出する．キーワードにはプライマリキーワードとセカンダリキーワードの 2 種類がある．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，質問文中の主題にあたる名詞などが該当する．残りのキーワードは全てセカンダリキーワードとする．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．検索されたウェブページから解答リストとなる表を抽出し，表の抽出に成功すればそれをユーザに提示する．表の抽出に失敗した場合は，従来のテキスト解析に基いて解答群を生成する手法を用い，生成された解答群をユーザに提示する．

解答リストを含む表を抽出する手法の詳細は以下の通りである．まずはじめに table タグで定義されている表を検出する．次に，表の 1 行目または 1 列目にあるセルとプライマ

リキーワードが一致するかを調べ，解答と関連のある属性を持つ表を抽出する．次に，すべてのセカンダリキーワードが (1) ウェブページの title タグの中，(2) 表のキャプション，(3) 表の前にある 3 つのセグメント，のいずれかに存在するかを調べ，存在しない場合はその表は質問のトピックと関連がないとみなして候補から除外する．最後に，プライマリキーワードが出現したセルと同じ行または列のセルが解答を含むかを調べる．表の 1 行または 1 列において，各セル内のテキストの固有表現タグと質問の解答タイプが一致している割合をしらべ，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．

次に，テキスト解析による解答群の生成手法の概要について述べる．まず，検索されたウェブページを HTML タグを用いてセグメント単位に分割し，キーワードを全て含むセグメントなど，解答候補が現れそうなセグメントを検索する．次に，固有表現タグや構文パターンを用いて解答候補を抽出する．抽出された解答候補の周辺にあるキーワードについて，キーワードと関連が高い語やキーワードと係り受け関係にある語を限定表現として抽出する．限定表現とは，曖昧なキーワードの意味を限定する表現のことで，冒頭に挙げた例では「サッカー」,「スキー」,「ラグビー」がそれにあたる．この段階で，(解答候補，キーワード，限定表現) といった 3 つ組が複数得られる．これら 3 つ組の集合から，キーワードが共通でかつ限定表現が何らかの共通属性を持つ解答候補をまとめ，解答群とする．一般に解答群は複数生成されるので，(1) 解答群の限定表現や解答の異なり数，(2) 限定表現の共通属性のタイプ，(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つける．最大のスコアをもつ解答群をユーザに提示する解答群とする．

本手法の評価を行なうために，曖昧な質問 30 個に対して，まずウェブページから表を抽出し，それに失敗した場合は解答群を生成するという方式で解答リストを出力する実験を行なった．その結果，56%の質問に対して，ウェブページから抽出した複数の表の中のいずれかか，テキスト解析によって生成されかつスコアが最大の解答群が正しい解答リストであった．また，83%の質問に対して，表の中に正解があるか，スコアの 10 位以内の解答群の中に正解が含まれていた．これらのケースでは，30 問のうち 9 問についてはウェブページから表を抽出し，残りの質問についてはテキスト解析によって生成された解答群を出力した．また，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は増えた．このことから，解答リストを得るために 2 つの手法を併用する提案手法は有効であることがわかった．