対話型質問応答システムにおける問い返し文の生成に関する研究

坂本 篤史 (510041)

北陸先端科学技術大学院大学 情報科学研究科

2007 年 2 月 8 日

キーワード: 質問応答, 対話, 情報抽出, 文生成, 語の意味の曖昧性.

本論文は曖昧な質問に適切に対応するオープンドメインな対話型質問応答システムについて述べる. 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のことである. 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ラグビー, バレーボール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. こうした質問に対して, システムが「どんなスポーツのワールドカップですか」などと問い返しを行えば, ユーザの返答によって適切な解答を取り出すことができる. 本論文は, 上記の対話型質問応答システムのうち, 質問の曖昧性検出と問い返し文生成の手法について述べる.

曖昧性検出までの本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, 文書検索を行い, 解答候補を抽出する. 次に, 解答候補が抽出された文書からキーワードの限定表現をパタンマッチによって抽出する. 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, 先ほどの例の「サッカー」,「ラグビー」,「バレーボール」がそれにあたる. 限定表現は, キーワードと同一段落中にある名詞の中でキーワードと関連が高い語, キーワードと係り受け関係にある名詞, 文書の先頭にある名詞などを抽出する. この段階で (a_i, k_j, s_k) という 3 つ組が複数得られる. ここで a_i は解答候補,k_j は質問文中のキーワード,s_k は k_j の限定表現である. これら 3 つ組の集合から, キーワードが共通でかつ s_k が何らかの共通属性 $attr$ を持つ部分集合を発見し, 解答群とする.$attr$ の例としては,「末尾 N 文字」(s_k の末尾 N 文字が共通),「意味クラス」(s_k の意味クラスが共通),「抽出パタン」(s_k の抽出パタンが共通) などがある. 解答群は $AG(k, attr) = \{(s_i, a_i)\}$ という形式で表現される. 例えば, 先ほどの「ワールドカップで優勝した国はどこですか」という質問に対して,AG(ワールドカップ, 意味クラス：スポーツ) $= \{$(サッカー, イタリア), (ラグビー, ニュージーランド), (バレーボール, キューバ)$\}$ といった解答群が得られる. 限定表現「サッカー」,「ラグビー」,「バレーボール」にはそれぞれ解答候補「イタリア」,「ニュー

ジーランド」,「キューバ」が対応し, またこれらの限定表現が「スポーツ」の意味クラスに属すことが共通である. 一般に解答群は複数生成されるので, キーワードの曖昧性を最も適切に表している解答群を選択するために解答群にスコアを与える. 解答群のスコアは, 限定表現と解答候補が 1 対 1 に対応しているか, 質問文中のキーワードと解答群中の限定表現がどれだけ意味的に関連しているか, 信頼度の高い抽出パタンによって抽出された限定表現をどれだけ含むか, などの観点で定義した.

解答群の作成はユーザの質問に含まれる曖昧なキーワード (k) を検出することに相当する. 同時に, 解答群の作成は, 問い返し文の内容を決定することにも該当する. なぜなら, 解答群の中の限定表現をユーザに指定してもらえば, 解答が 1 つに決まるからである.

次に, 解答群からユーザに対する問い返し文を生成する処理について述べる. 基本的には, ユーザからキーワードの限定表現を聞き出す質問をテンプレートによって生成する. テンプレートは 3 種類用意した.1 つ目は, 解答群の限定表現の数が 2 種類のとき, 二者択一の疑問文を生成するテンプレートである.2 つ目は, 問い返し主題を含む問い返し文を生成するテンプレートである. 問い返し主題とは,「どんなスポーツ」というように, 解答群の限定表現全体を指す語 (スポーツ) に疑問詞 (どんな) をつけた表現である.3 つ目は問い返し主題を含まない文を生成するテンプレートである. 解答群によっては問い返し主題を生成できない場合があり, このときは曖昧なキーワードだけを使って問い返し文を生成する. 本研究では合計 23 個の生成テンプレートを用意した.

テンプレートから複数の問い返し文の候補を生成し,n-gram の頻度によるスコア付けを行った. すなわち, 問い返し文の部分文字列 (n-gram) の新聞記事中における出現頻度が多いものに高いスコアを与えることで, 日本語として自然な文を選択した. ただし, 疑問詞は新聞記事にあまり現れないので, 疑問詞を含む問い返し主題は Web の検索エンジンのヒット数によって複数の候補の中から最適なものを 1 つ選択した. また, 問い返し文における限定表現とキーワードの出現順序も n-gram の頻度によるスコアでは決定できないので, 解答を取り出した根拠文における両者の位置関係を参照して決めた.

曖昧な質問 50 問に対して問い返し文を生成する実験を行った. その結果, 解答群の正解率が 64%, 問い返し文生成の正解率が 76% であった. これにより提案手法の有効性が確認された.