

Title	曖昧な質問に対応する対話的質問応答システムに関する研究
Author(s)	徳江, 英範
Citation	
Issue Date	2005-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1911
Rights	
Description	Supervisor:白井 清昭, 情報科学研究科, 修士

曖昧な質問に対応する対話的質問応答システムに関する研究

徳江 英範 (310073)

北陸先端科学技術大学院大学 情報科学研究科

2005 年 2 月 10 日

キーワード: 質問応答, ユーザとの対話, 曖昧な質問, 情報抽出.

本論文は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. このシステムは, ユーザの質問が曖昧であるときに, その曖昧性を解消するためにユーザに問い返しを行い, それに対するユーザの返答に基づいて最適な回答を選択する. ここでの曖昧な質問とは, 質問中の単語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. 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, これに対する解答を一意に決めることはできない. このような曖昧な質問が入力されたとき, 「どの競技のワールドカップですか」といった問い合わせをシステムからユーザに行い, これに対するユーザからの返答によって適切な解答を決める. 本研究では, このような対話型質問応答システムの要素技術として, ユーザの質問の曖昧性を検出する手法について主に述べる. 提案手法の概要は以下の通りである. 先の例で, 「サッカーのワールドカップの優勝国はブラジル」と「ラグビーのワールドカップの優勝国はイギリス」のように, 「ブラジル」「イギリス」と2つの解答候補が得られたとする. ここで, キーワードの意味を限定するような表現 (限定表現) に着目する. 例えば, 「ワールドカップ」に連体修飾する「サッカーの」「ラグビーの」という句は「ワールドカップ」の意味を限定する表現とみなせる. このように, 質問文に含まれるキーワードについて, そのキーワードの限定表現を抽出し, 同じキーワードに対して解答候補毎に異なる限定表現が存在すれば, そのキーワードは曖昧であるとみなす.

本システムの処理における曖昧性検出までの流れは以下の通りである. まず, ユーザからの質問文からキーワード, 解答タイプ, キーワードタイプを抽出し, キーワードにマッチする文書を抽出する. 抽出した文書から解答候補を抽出する際, 本研究では, 形態素情報が解答タイプにマッチしている, 主キーワードの近傍にあるという2つの条件を満たす名詞を解答候補として抽出した. このようにして抽出した解答候補にスコアを付け優先順位を決める. スコアの式を以下に示す.

$$S_{ans} = w_{pat} \times S_{pat} + w_{mor} \times S_{mor} + w_{dis} \times S_{dis}$$

S_{pat} はキーワードタイプによるスコア, S_{mor} は解答の形態素情報によるスコア, S_{dis} はキーワードと解答候補の距離によるスコアである. w_x はそれぞれのスコアの重みを示す.

次に、質問の曖昧性の検出を行うために、解答候補が抽出された文書からキーワードに係る表現を抽出する。抽出する限定表現は、キーワードの「連体修飾句」、「格助詞デ格」に係る名詞、キーワードが複合名詞を作る場合その「直前の単語」、「直後の単語」、そして「日付表現」の5つである。例えば、キーワードが「金メダリスト」のとき、「柔道の金メダリスト」という文からキーワードに連体修飾する「柔道」という単語が限定表現として抽出される。そして、キーワードと限定表現抽出の種類毎に限定表現のグループを作り、そのグループ内で個々の解答候補が異なる限定表現を持つかを調べることによって曖昧性の検出を行う。しかし、このグループ内には、曖昧性を検出するには不要な限定表現も含まれることがよくある。そこで、本研究では、属性という概念を導入して不要な限定表現を取り除くことを試みた。属性とは限定表現が持つ特徴のことを指す。本研究では、属性として「数詞＋接尾語」「かぎ括弧」「意味クラス」「末尾N文字」「日付表現」の5種類を扱う。例として、「5代目」「3代目」「水戸」という限定表現のグループがあった場合、「数詞＋接尾語」の属性では、「5代目」「3代目」が抽出され、「水戸」という限定表現は属性を持たない限定表現とみなし削除する。このように、初期の限定表現のグループ内で限定表現同士が共通の属性を持っていれば、それを改めて1つのグループとみなし、グループ内で属性を持たない限定表現を取り除く。次に、このような一限定表現のグループは複数得られるため、これらに優先順位をつけるためのスコア付けを行い、問い返しに最適なグループを1つ選択する。スコアは、多くの解答候補で限定表現が現れているか否か、属性を持つ限定表現の頻度、異なる解答候補に対して同じ限定表現が得られているかどうか、などを考慮して決める。本手法の評価を行うために、曖昧性を含む質問33個に対して上記の手法で曖昧性の検出を行う実験を行った。その結果、曖昧な質問であると判断された質問は全体の80%程度であり、そのうち、適切な属性で限定表現が抽出された割合は約80%であった。