

Title	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
Author(s)	坂本, 篤史
Citation	
Issue Date	2007-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/3574
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 修士

修 士 論 文

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

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

坂本 篤史

2007 年 3 月

修 士 論 文

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

指導教官 白井 清昭

審査委員主査 白井 清昭 助教授
審査委員 島津 明 教授
審査委員 鳥澤 健太郎 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

510041 坂本 篤史

提出年月: 2007 年 2 月

概要

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おける問い返し文生成について述べる。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。そこで、本システムは、このようにユーザの質問が曖昧である場合に、まず質問文のキーワードについて、その意味を限定する表現(限定表現)を抽出する。先ほどの例では、限定表現は「サッカー」、「ラグビー」となる。次に、「サッカー」の解答として「イタリア」が、「ラグビー」の解答として「ニュージーランド」が存在すれば、限定表現とその解答の組(「サッカー:イタリア」、「ラグビー:ニュージーランド」)を作る。さらに、「サッカー」と「ラグビー」はどちらも「スポーツ」の意味クラスに属することが共通であるので、この2つの組をまとめて解答群を作成する。このようにして、複数の解答群を生成し、解答群にスコアをつける。

最もよいスコアの解答群に対して、問い返し文の候補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種類するとき、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。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生成の実験を行い、54%の質問に対して適切な問い返し文が生成されることを確認した。

目次

第1章	序論	1
1.1	研究の背景	1
1.2	研究の目的	1
1.3	本論文の構成	2
第2章	関連研究	3
2.1	一般的な質問応答システム	3
2.2	クローズドドメインでの対話型質問応答システムに関する研究	4
2.3	オープンドメインでの対話型質問応答システムに関する研究	5
2.4	本研究の特色	6
第3章	提案システム	7
3.1	質問応答システムの概要	7
3.2	モジュール	7
3.2.1	質問文解析	8
3.2.2	文書検索	9
3.2.3	解答候補抽出	9
3.2.4	解答群作成	12
3.2.5	問い返し文の表層表現	12
3.2.6	解答の提示	13
第4章	解答群の作成	14
4.1	限定表現候補の抽出	14
4.2	属性の付与	16
4.3	解答群の作成	18
4.4	解答群に対するスコア付け	22
4.5	本手法と松本らの手法との比較	24
第5章	問い返し文の生成	27
5.1	処理の流れ	27
5.2	問い返し主題の生成	29
5.3	テンプレートの適用条件	34

5.4	n-gram の頻度によるスコアリング	35
5.5	自然な問い返し文を生成するための工夫	37
第 6 章	評価実験	40
6.1	実験方法	40
6.2	考察	42
6.2.1	解答群作成の考察	42
6.2.2	問い返し文の生成の考察	46
6.2.3	解答群の選択処理と問い返し文の生成処理の融合	51
第 7 章	結論	52
付 録 A	実験に用いた質問一覧	57

目 次

1.1	提案システムによる対話例	1
4.1	記事の先頭例	16
4.2	限定表現抽出の例	16
4.3	AG(ワールドカップ,E3:ー女子)の解答群	19
4.4	(ワールドカップ,N:数+回)の解答群	19
4.5	(優勝,T:898f)の解答群	20
4.6	解答群を構成する要素が1つの例	20
4.7	解答群を構成する限定表現が全て同じ例	20
5.1	序数詞一覧	28
5.2	テンプレートCによる生成例	29
5.3	問い返し文の生成の処理の流れ	30
5.4	表 5.10 の解答を含む根拠文	35
5.5	どちらが日本語として正しい文か?	37
6.1	適切な解答群の抽出例	42
6.2	図 6.1 の記事	43
6.3	パターン「記事の先頭」によって抽出された限定表現を含む解答群の例	44
6.4	図 6.3 の記事	45
6.5	日本が入っていない解答群	45
6.6	問い返し文の成功例	47
6.7	Web 検索のヒット数の失敗例 (解答群)	48
6.8	キーワードと限定表現の位置の失敗例 (解答群)	49
6.9	キーワードと限定表現の位置の失敗例 (問い返し文生成,N-gram 選択)	50
6.10	他の失敗例 (解答群)	50
6.11	他の失敗例 (問い返し文選択)	51

表 目 次

3.1	解答群の例	7
3.2	解答タイプ	9
3.3	キーワードタイプ	9
3.4	解答候補が満たすべき条件	10
3.5	解答候補抽出パターンとスコア	11
3.6	品詞情報のスコア	11
4.1	表 4.2 の記事からの限定表現抽出結果	17
4.2	表記のゆれの吸収の例 (数量表現 + 接尾語 (N))	17
4.3	各限定表現の属性	19
4.4	限定表現の表記のゆれの吸収の例	21
4.5	解答の表記のゆれの吸収の例 (解答タイプが人名)	21
4.6	解答の表記のゆれの吸収の例 (解答タイプが人名以外)	22
4.7	属性のスコア	23
4.8	限定表現の抽出パタンのスコア	24
4.9	「サリン事件が起こったのはいつですか」についての解答群	25
5.1	問い返し文テンプレート A	27
5.2	問い返し文テンプレート B	28
5.3	問い返し文テンプレート C	29
5.4	共通末尾の例	31
5.5	意味クラス + 共通末尾の例	31
5.6	解答群「芥川賞を受賞したのはだれですか」	32
5.7	接尾語の末尾の部分文字列を助数詞とする例	32
5.8	解答群「野球の新人王を受賞したのはだれですか」	33
5.9	解答群「世界選手権で優勝した国はどこですか」	33
5.10	解答群「世界選手権で優勝した国はどこですか」	35
5.11	予備実験での n -gram の頻度に対する重み	37
5.12	予備実験の結果 (28 問中)	37
5.13	予備実験の質問一覧	38
5.14	「世界チャンピオン」 k と「何級」 $wh+s$ の位置の違い	39

6.1	解答群作成の実験結果	41
6.2	問い返し文の生成の実験結果	41
6.3	本手法の実験結果	41
6.4	限定表現とキーワードの Dice 係数	44
6.5	スノーボードの種目の Dice 係数	46
6.6	Web 検索のヒット数の失敗例 (Web 検索のヒット数)	48
6.7	問い返し文の生成の失敗例	49
6.8	他の失敗例 (Web 検索のヒット数)	50
7.1	解答群中の限定表現の選定	54
A.1	オープンテストの質問一覧 1	57
A.2	オープンテストの質問一覧 2	58
A.3	オープンテストの質問一覧 3	59
A.4	オープンテストの質問一覧 4	60
A.5	オープンテストの質問設定 1	60
A.6	オープンテストの質問設定 2	61

第1章 序論

1.1 研究の背景

近年、情報技術の発達により、ユーザの知りたい情報が容易に手に入るようになった。しかし、情報が増大しすぎたために、正しい情報を手に入れるのに時間がかかったり、手に入れた情報を活用できなかったりしている。そのような状況においても質問応答システムは、ユーザの質問に対する解答のみを表示することで情報を有意義に活用できる。1990年代から、電子化された大量のテキストが利用可能になってきたこともあり、テキストをデータベースとした質問応答システムの研究が盛んに行われ始めた。現在の質問応答システムの主流は、1つの質問に対して1つの解答を返す一問一答型のシステムである。しかし、質問によっては単語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場合に解答を返せないことが問題となっている。例えば、「アカデミー賞の受賞者はだれですか」という質問は、アカデミー賞には主演女優賞や監督賞など様々な賞があるために解答を1つに絞ることができない。

1.2 研究の目的

本研究では、曖昧な質問に対して、システムからユーザに問い返しを行うことで問題を解決する。例を図 1.1 に示す。曖昧な質問「アカデミー賞の受賞者はだれですか」に対して「アカデミー賞の何賞ですか」と問い返し、これに対し、ユーザが「監督賞です」と答えれば解答を1つに絞り込める。

ユーザ：「アカデミー賞の受賞者はだれですか」 システム：「アカデミー賞の何賞ですか」 ユーザ：「監督賞です」 システム：「スティーヴン・スピルバーグです」
--

図 1.1: 提案システムによる対話例

本研究では、このような対話型質問応答システムにおいて、以下の2つのことを行うのが目的である。

- 質問の曖昧性の検出

松本らの研究 [1] によるリスト型質問応答システムの手法を基盤として, 質問の曖昧性検出を行う. これは, 問い返し文の内容を決定する処理に相当する. 松本らの手法は, 曖昧性の検出の精度が低いため, 彼らの検出手法を洗練する.

- 問い返し文の表層表現の生成

質問の曖昧性検出で得た情報を使い, ユーザーへの問い返し文の具体的な表層表現の生成を行う. 問い返し文は, 曖昧なキーワードの意味をユーザに問う文であり, なおかつ日本語として自然な文を生成する.

1.3 本論文の構成

本論文の構成は以下の通りである.

2 章では, 関連研究や本研究の特色について述べる.

3 章では, 本研究で提案する対話型質問応答システムの概要について述べる.

4 章では, 問い返し文の内容を決定するための手法を述べる.

5 章では, 問い返し文の表層表現を生成するための手法を述べる.

6 章では, 提案手法の評価実験とその考察について述べる.

7 章では, 本研究のまとめ, 及び今後の展望について述べる.

第2章 関連研究

2.1 一般的な質問応答システム

TREC や QAC などの評価型ワークショップの参加システムに代表されるように、ユーザの質問に対してシステムが解答を返すだけで処理が終了する一問一答型のシステムが現在の主流である。

Abraham Ittycheriah らは、TREC-9 における質問応答システムを構築した [2]。まず、ユーザの質問を解析し、maximum entropy model によって MUC の固有表現抽出カテゴリーを基準とした解答タイプに分類する。また、ユーザの質問からクエリー展開し解答がある文書を検索する。そのためには、文書に固有表現抽出を行わなくてはならない。maximum entropy model を訓練し MUC の固有表現タグを付与した。そして、抽出された文書中から解答タイプに適合する解答を抽出した。

石塚らは、グラフ構造を用いた距離尺度を導入した質問応答システムを構築した [3]。質問文をあらかじめ 4 つのタイプに分けておく。タイプ 1 は「木星は何個の衛星を持っていますか」など「何 + 単位」もしくは「何 + 接尾語」という質問のタイプで、解答候補はこのような単位を持った表現に限定される。タイプ 2 は「だれ」「どこ」「いつ」を尋ねる質問のタイプで、解答候補は「人名」「地名、会社名、機関名」「時間、日付」に限定される。タイプ 3 は「どのくらいですか」と尋ねる質問のタイプで、解答候補は数量表現に限定される。タイプ 1 と異なる点は単位がわからないことである。タイプ 4 は解答候補の属性や表現を限定できない「マカオはポルトガル語でどのように表しますか」などの質問のタイプである。処理の流れは、質問文を形態素解析し、キーワードを抜き出す。そのキーワードを使って文書検索する。もし、文書が見つからなければ、キーワードを減らして文書を検索する。これを文書が見つかるまで繰り返す。次に見つけた文書のキーワードを含む文から質問タイプに合う解答候補を探す。もし解答候補が見つからなければ、その文の前後の文を加えて解答候補を抽出する。複数得られた解答候補に対して順位付けを行う。まず、解答候補を抽出した文書を係り受け解析する。係り受け関係に従い、各文節をノードとするグラフ構造を作成する。グラフ内のノードに対し、キーワードかそうでないか分割する。有向グラフを無向グラフに変換し、隣接するノード間のコストを決める。Dijkstra のアルゴリズムに従って、解答候補とキーワードの最短距離を算出する。そして、ある解答候補とすべてのキーワードとの距離の和を、その解答候補のスコアとして順位付けを行う。

本研究は対話を通じてユーザの質問に対する適切な解答を提示するシステムの構築を目指す点でこれらの一問一答型の質問応答システムに関する研究と異なる。

2.2 クローズドドメインでの対話型質問応答システムに関する研究

対話型質問応答システムとは、ユーザとシステムが交互に質問と返答を繰り返すシステムである。ユーザの質問が曖昧な場合は、システムから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解答を選択し、曖昧性を解消する。

清田らは、Windows 環境でのパーソナルコンピュータを利用している人を対象にした対話型質問応答システム「ダイアログナビ」を構築した [4]。マイクロソフトがすでに一般に公開しているテキスト知識ベースと対話カードを用いて問い返しを行う。対話カードとは、曖昧なユーザ質問文のうち典型的なものに対して、どのような聞き返しを行うかを記述したカードのことである。まずユーザの質問が対話カードの文と一致するか判断する。一致する場合は、対話カードにしたがった問い返しを行う。一致しない場合は、ユーザの質問とテキスト知識ベース中の文と一致した複数の文から状況説明文を抽出し、ユーザに提示することで曖昧性を解消する。状況説明文とは、テキスト知識ベース中の文と一致しなかった箇所のことである。例えば、「音が出ない」という質問に対し、テキスト知識ベースに「Windows サウンド (.WAV) ファイルを再生時に、音が出ない」という文と「コントロールパネルの [サウンド] から CHIMES.WAV ファイルをテストした場合、ボリューム設定に関わらず、音は出ません。」という文が抽出された場合に、ユーザの質問「音が出ない」という箇所を除いた「Windows サウンド (.WAV) ファイルを再生時に」と「コントロールパネルの [サウンド] から CHIMES WAV ファイルをテストした場合、ボリューム設定に関わらず」の部分ユーザに提示して解答を選択させる手法である。

黒橋らは、京都大学総合情報メディアセンターが提供する計算機システム、アプリケーション、ソフトウェアについて、利用者の質問に答える対話的ヘルプシステムを構築した [5]。ユーザの質問を KNP によって構文解析し、発話タイプに分類する。発話タイプには、事実を問う What 型や方法を問う How 型、症状を提示しその対処を求める Symptom 型、メディアセンターに対して要求する Request 型、以前の質問の修正・追加をする Addition 型、システムの質問に対して返答する Answer 型の 6 つがある。システムは、What 型、How 型、Symptom 型の質問に回答し、さらに Addition 型および Answer 型の発話を文脈に応じて適切に解釈する枠組みを持っている。Request 型については対象外としている。次に、ユーザの質問を解析した情報と知識ベースとのマッチングを行い、最も類似度の高い部分を見つけて、それに対応する回答を表示する。知識ベースには、見出し語とその説明文という辞書のような形式で与えている。ユーザとの対話には、未知語の聞き返し、文脈補完処理による文脈に依存した入力文の解釈、曖昧な質問に対する聞き返し、挨拶に対しての返事が行える。

Denecke らは、Hierarchical Tree Kernel を用いて問い返しを行う質問応答システムを構築した [6]。知識ベースにはユーザとウィザードの発話が交互に現れる対話コーパスを用意した。ユーザから質問された場合に、その対話コーパスからもっとも類似する発話を Hierarchical Tree Kernel によって抽出する。そして、最も類似した発話の次の発話に対して、現在の質問に合うように、文脈と新聞記事から単語を置換する。すると、複数の問い返し文が得られる

ので、複数の問い返し文同士を Hierarchical Tree Kernell を使ってスコア付けをする。最も高いスコアを得た問い返し文をユーザに提示する。なお、この研究の対象ドメインは、対話コーパスの例から観光案内だと推測される。

本研究はオープンドメインな対話型質問応答システムの構築を目指すという点でこれらの研究と異なる。

2.3 オープンドメインでの対話型質問応答システムに関する研究

質問の曖昧性を考慮した対話型質問応答システムに関する研究のうち、徳江らは、オープンドメインな対話型質問応答システムの研究を行った [7]。質問を解析したときに得られるキーワードの意味を限定する表現 (限定表現) を抽出し、同じ解答候補ごとに異なる限定表現が存在したときに、そのキーワードが曖昧であるとみなしている。限定表現を抽出するために、キーワードにかかる限定表現のパターンを記述しておき、そのパターンに一致した限定表現を抽出する。例えば、「ワールドカップ」がキーワードのとき、「サッカーのワールドカップ」という文から、パターンマッチによって「サッカー」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の属性ごとにグループを作り、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検出を行う。ここで属性とは「数量表現 + 接尾語」「かぎ括弧」「意味クラス」「末尾 N 文字」「日付表現」を表す。例えば、「サッカー」「ラグビー」「バトミントン」は「球技」という同じ「意味クラス」に属する。そして得られたいくつかの限定表現のグループ (解答群) をスコア付けし、問い返しに最適なグループを 1 つ選択している。しかし、パターンマッチによる限定表現の抽出数が少ないために適切な解答群が得られないことが多い。また、具体的な問い返し文の生成手法については述べられていない。

松本らは、曖昧な質問に対して解答群をリスト表示する質問応答システムの構築を行った [1]。例えば、「アカデミー賞の受賞者はだれですか」という質問の場合、アカデミー賞には主演女優賞、助演女優賞、監督賞など様々な賞があり、それぞれの限定表現に対して解答が異なっている。そこで、解答と限定表現の組を作り、それをリストとして表示する。先ほどの例は「グウィネス・パルトロー:主演女優賞」「ジュディ・デンチ:助演女優賞」「スティーヴン・スピルバーグ:監督賞」のようなリストになる (: の前は解答、後は限定表現を表す)。ユーザは限定表現によって解答の違いを理解することができ、知りたい解答をリストの中から見つけることができる。これに対し、本研究はユーザに問い返しを行うことで質問の曖昧性を解消する点で松本の手法と異なる。しかし、上記のようなリストを作成することは本研究における問い返し文生成の内容を決定する部分に利用できる。松本らが作成した解答リストでは、解答と限定表現の組がキーワードと属性でまとまっている。ユーザに限定表現を特定してもらえば適切な解答を返せるので、本研究はその限定表現を特定してもらえらるような問い返し文の表層表現を生成すればよいことになる。

2.4 本研究の特色

これまでに挙げた関連研究と本研究の違いは以下のとおりである。

- 曖昧な質問に対応する

2.1 節で挙げた質問応答システムと違い、ユーザの質問が曖昧であった場合に対応するという点が異なる。曖昧であるとは、質問文中のキーワードに明確でない点が存在し、キーワードの意味が一意に決められず、解答が絞り込めないということを指す。「アテネオリンピックのマラソンの金メダリストはだれですか」という質問の場合、マラソンには男子、女子があるため解答が1つに絞り込めない。この場合、マラソンというキーワードは曖昧であると考えられ、このキーワードを使い「男子マラソンですか、女子マラソンですか」といった問い返し文を生成することで曖昧性検出を行う。

- 曖昧性検出の精度を高める

徳江らや松本らの研究における解答群の作成は質問文に含まれる曖昧性を検出することに相当する。本研究は、限定表現抽出パターンや解答群を作成するための共通属性の数を増やすことで曖昧性検出の精度を高める。

- 問い返し文を生成する

問い返し文の内容を決定した後、その表層表現を生成する。解答群中の限定表現と曖昧なキーワードを問い返し文生成テンプレートに当てはめ、問い返し文を複数生成する。その中から最も日本語として適切なものを n-gram によって1つ決定し、ユーザに提示する。

本研究の特色は、オープンドメインな対話型質問応答システムを構築することにある。松本の手法を基に解答群を作成し、ユーザが求めている解答に対する限定表現を答えさせるような問い返し文を生成する。さらに、解答群を作成する松本の手法における問題点の克服にも取り組む。

第3章 提案システム

3.1 質問応答システムの概要

本研究は曖昧な質問に対応した質問応答システムの構築を行った。質問の中のキーワードの意味が曖昧な場合、その曖昧性を検出し、問い返し文の内容を決め、問い返し文の表層表現をユーザに提示し、それに対してのユーザの返答を受け取り、解答をユーザに返す。例えば「アカデミー賞を受賞したのはだれですか」という質問がある。アカデミー賞には「主演女優賞」「助演女優賞」「監督賞」など様々な賞があるために、解答が1つに絞ることができない。そこで問い返し文を生成し、ユーザにキーワードの意味を限定してもらう。本研究では、この「主演女優賞」「助演女優賞」「監督賞」といった表現をキーワード（この場合は「アカデミー賞」がキーワード）の意味を限定することから限定表現と呼ぶ。限定表現とその解答で組を作り、表3.1のように限定表現の共通する属性（この場合は「賞」が共通属性）で解答群を作る。ここで仮に、ユーザに限定表現「監督賞」を指定してもらえば、その解答「スティーヴン・スピルバーグ」を返すことができる。つまり、ユーザから限定表現を聞き出すようにシステムからユーザへ問い返しを行えば、よいことになる。よって、解答群を作ることは問い返し文の内容を決定する処理に相当する。

表 3.1: 解答群の例

質問：アカデミー賞を受賞したのはだれですか	
限定表現	解答
主演女優賞	グウィネス・パルトロー
助演女優賞	ジュディ・デンチ
監督賞	スティーヴン・スピルバーグ

次に、図3.1の解答群を使い、問い返し文の表層表現を決定する。限定表現に共通している賞と疑問詞を組み合わせ「どんな賞」という表現とキーワード「アカデミー賞」から「アカデミー賞のどんな賞ですか」と問い返し文の表層表現を生成する。

3.2 モジュール

本システムは、以下の順で処理を進める。

1. 質問文解析
2. 文書検索
3. 解答候補抽出
4. 解答群作成
5. 問い返し文の表層表現
6. 解答の提示

なお、質問文解析、文書検索、解答候補抽出で述べる処理は、松本らのシステム [1] の処理と同じである。以下、これらの処理の概要を述べる。

3.2.1 質問文解析

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。しかし、現状では質問文の解析を自動で行ってない。本研究の主題の1つは質問文の曖昧性検出なので、質問文解析の段階で不適切な解析結果が含まれるのは適当ではないからである。よって、適切な質問解析ができているものとして人手で質問文を解析した。

- キーワード

キーワードとは、入力された質問文の中から解答候補の手がかりとなる名詞である。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がある。

- － プライマリキーワード

文書検索かつ解答候補抽出のための中心となるキーワードである。検索された記事の中の、プライマリキーワードがある文から解答候補を抽出する。質問文のキーワードでプライマリキーワードは1つである。

- － セカンダリキーワード

文書を検索するためのキーワードであり、解答候補抽出には使われない。セカンダリキーワードは1つ以上である。

- 解答タイプ

解答タイプとは、質問文が解答として求める情報の種類のことである。解答タイプはIREX[8]の固有表現タグに準じて、「人名」、「国名」、「地域名」、「組織名」、「時間」とIREXの固有表現タグに対応するものがない「その他」の6種類を設定した。解答タイプの詳細を表3.2に示す。

- キーワードタイプ

キーワードタイプとは、プライマリキーワードと解答候補の間に成り立つ関係を示すタイプである。表3.3にキーワードタイプの一覧を示す。

表 3.2: 解答タイプ

解答タイプの種類	解答タイプの概要
人名 [per]	「だれ」を尋ねるタイプ
国名 [na]	「どこ」を尋ねるタイプで国を示す
地域名 [loc]	「どこ」を尋ねるタイプで地域の名称を示す
組織名 [org]	「どこ」を尋ねるタイプで会社などの組織を示す
時間 [time]	「いつ」を尋ねるタイプで時間に関わることを示す
その他 [ot]	上記の解答タイプ以外

表 3.3: キーワードタイプ

名前	パターン	キーワードと解答の関係
hyponym	< 解答候補 > は < キーワード > だ	上位-下位関係
agent	< 解答候補 > が < キーワード > する	動作主の関係
other	その他	上記以外

質問文の解析例を以下に挙げる。「芥川賞を受賞したのはだれですか」という質問文を解析する。形態素情報から「芥川賞」、「受賞」というキーワードが抽出され、プライマリキーワードは「受賞」、セカンダリキーワードは「芥川賞」となる。「だれ」という情報から「人名」が解答タイプとなる。「受賞したのはだれ」というパターンから「受賞した」と「だれ」の間には動作主の関係が成り立っているため、キーワードのタイプが「agent」と決まる。

3.2.2 文書検索

質問文解析で得られたキーワードで検索し、全てのキーワードが出現する文書を抽出する。データベースとして毎日新聞 1991 年から 2003 年までを用いた。検索年は 1991 年から 2003 年の任意の期間を指定できる。また動詞、形容詞などの活用形に対応するために転置インデックスを作成した。例えば「走る」という動詞をキーワードとした場合、文書に「走った」と表記されていても抽出できるようにした。これによって文書検索の再現率が向上する。

3.2.3 解答候補抽出

文書検索で得られた文書を対象に解答候補を抽出する。抽出された記事を入力とし、解答候補と解答候補のスコアが出力される。以下に解答候補抽出の処理の流れを示す。

1. 抽出された記事の解析

形態素解析, 構文解析をする. 形態素解析には茶筌 [9], 構文解析には南瓜 [10] を使った.

2. 解答候補の抽出

抽出された文章から, 以下の条件を満たす名詞を解答候補として抽出する.

- 形態素情報が解答タイプの条件を満たす

ここで必要とする形態素情報とは, 固有表現タグ, 品詞タグ, カタカナ文字列の3種類である. 例えば, 解答タイプが「人名」の場合に, 固有表現タグが<PERSON>か, 品詞が<名詞-固有名詞-人名-*>か, カタカナ文字列であるかという条件のいずれかを満たす名詞であれば解答候補とする. 表 3.4 に解答タイプごとに満たすべき条件を示す.

表 3.4: 解答候補が満たすべき条件

解答タイプ	固有表現タグ	品詞タグ	カタカナ
人名	<PERSON>	名詞-固有名詞-人名-*	カタカナ文字列
国	<NATION>	名詞-固有名詞-地域-国	カタカナ文字列
地名	<LOCATION>	名詞-固有名詞-地域-一般	
組織名	<ORGINIZATION>	名詞-固有名詞-組織	
時間	<DATE>		
その他		名詞全般	

- プライマリキーワードの近傍にある

プライマリキーワードと解答候補の構文パターンを作り, 構文パターンに適合する名詞を解答候補として抽出する. 表 3.5 に構文パターンを示す. 「→」は文節の係り先を示す. 3.2.1 項で述べたキーワードタイプによって用いる構文パターンが異なる. 表 3.5 の「近傍の名詞」はキーワードの近傍にある名詞を抽出するパターンを表す. 解答候補のスコアについては後で述べる.

- 解答候補のスコア

解答候補がどれだけ解答としてふさわしいか解答候補にスコアを付けた. スコア付けは以下の3つの要素によって決まる.

- 構文パターン

プライマリキーワードと解答候補に関するスコア

- 品詞パターン

解答候補の品詞情報に関するスコア

- 距離スコア

解答候補と各キーワードの距離に関するスコア

解答スコアの算出式を式 (3.1) に示す.

表 3.5: 解答候補抽出パターンとスコア

構文パターン	適用されるキーワードタイプ	解答候補のスコア
<解答候補>ハ→<キーワード>だ	hyponym	1
<キーワード>ハ→<解答候補>だ	hyponym	1
<キーワード>の→<解答候補>	hyponym	0.8
<キーワード>, <解答候補>	hyponym	0.8
<キーワード>である<解答候補>	hyponym	0.8
<解答候補>ガ→<キーワード>する	agent	1
上記にないパターン「近傍の名詞」	hyponym,agent,other	0.1

$$S_{ans} = w_{pat} \times S_{pat} + w_{mor} \times S_{mor} + w_{dis} \times S_{dis} \quad (3.1)$$

構文パターンにおけるスコア ($w_{pat} \times S_{pat}$)

プライマリキーワードと解答候補の間に決められた係り受け関係が存在する場合に適用されるスコアである。スコアの詳細は表 3.5 で示した。また, w_{pat} はスコアに対する重みを表しており, 値は 0.4 である。

品詞パターンにおけるスコア ($w_{mor} \times S_{mor}$)

解答候補を含む文を茶筌, 南瓜によって形態素解析, 固有表現タグ付けを行い, 条件に合った品詞情報のスコアを解答候補に与える。解答候補の属性に対するスコアを表 3.6 に示す。また, w_{mor} はスコアに対する重みを表しており, 値は 0.1 である。

表 3.6: 品詞情報のスコア

属性	スコア
固有表現タグ	1
品詞	0.9
カタカナ	0.8

距離情報におけるスコア ($w_{dist} \times S_{dist}$)

ここでは各キーワードと解答候補によるスコアを計算する. 距離スコアの算出式を式 (3.2) に示す. このスコアは解答候補とキーワードの距離が近いほど, スコアが高くなるようにしたものである.

$$S_{dist} = \frac{1}{K} \sum_{i=1}^N \left\{ \frac{1}{dist(a_i, k_j) + 1} \times sent(a_i, k_j) \right\} \quad (3.2)$$

- K ...総キーワード数
- a_i ...解答候補
- k_j ...キーワード
- $dist(a_i, k_j)$... a_i と k_j の距離
-

$$sent(a_i, k_j) = \begin{cases} 0.2 & (a_i \text{と} k_j \text{が一文中に存在しない場合}) \\ 1 & (a_i \text{と} k_j \text{が一文中に存在する場合}) \end{cases}$$

ここで, 「 a_i と k_j の距離」とは, 解答候補とキーワードの間に存在する文字数を示す. キーワードが解答候補の前に存在する場合は, キーワードの末尾と解答候補の先頭の間の文字列の長さを計算する. キーワードが解答候補よりも後ろに存在する場合は, 解答候補の末尾からキーワードの先頭までの文字列の長さを計算する. キーワードと解答候補が隣接している場合は距離が0になり, 分母が0になるため, 距離に1を加える. また, キーワードと解答候補が句読点をまたいで出現している場合は, 話題が変わっている可能性があるため, それぞれのキーワードについて, 同一文中にキーワードと解答候補が存在しないときにはスコアを低くする. これは式 (3.2) の $sent(a_i, k_j)$ によって実現されている. また, w_{dist} はスコアに対する重みを表しており, 値は0.5である.

3.2.4 解答群作成

解答候補が得られた文書を対象にキーワードの意味を限定する表現をパターンマッチにより取り出し, 限定表現の候補とみなす. 抽出した限定表現の候補が持つ属性が共通する解答でグループを形成し, スコアリングによって最適な解答群を生成する. これは問い返し文の内容を決定するモジュールである. 詳しくは4章で述べる.

3.2.5 問い返し文の表層表現

最もスコアの高い解答群から問い返し文の表層表現を決定し, ユーザに対して問い返しを行う. 詳しくは5章で述べる.

3.2.6 解答の提示

以下の手続きでユーザが求めている解答を返すことができる。まず、システムが生成した問い返し文に対してユーザが答える。次に、ユーザの返答から解答群中の限定表現と一致するものを探し、一致した限定表現と対になっている解答をユーザに返す。

この処理は本研究ではまだ実装されていない。解答が一意に決まらない場合は再度ユーザに問い返すといった対話制御の処理も含めて、ユーザの返答から最適な解答を1つ選択する手法を確立することは今後の課題である。

第4章 解答群の作成

この章では、本研究の解答群の作成について述べる。本研究では質問文に含まれる曖昧なキーワードが複数存在する場合でも、曖昧なキーワードを1つ選び、そのキーワードの意味を限定する表現の共通属性で解答群を形成する。本研究ではキーワードの意味を限定する表現を限定表現と呼ぶ。この章で述べる処理は松本らの手法[1]を改善したものである。

4.1 限定表現候補の抽出

限定表現を抽出するために7つのパターンを用意した。

- 連体修飾 (s_{no})
助詞「の」を介してキーワードに連体修飾する句を限定表現として抽出する。例えば、「サッカーのワールドカップ」という表現があったとき(「ワールドカップ」がキーワード)、「ワールドカップ」の限定表現 s_{no} は「サッカー」となる。
- 直前の単語 (s_{prev})
キーワードの直前にあり、キーワードとともに複合名詞を構成する名詞を限定表現として抽出する。例えば、「第17回ワールドカップ」という表現があったとき(「ワールドカップ」がキーワード)、「ワールドカップ」の限定表現 s_{prev} は「第17回」となる。
- 直後の単語 (s_{succ})
キーワードの直後にあり、キーワードとともに複合名詞を構成する名詞を限定表現として抽出する。例えば、「ワールドカップ日韓大会」という表現があったとき(「ワールドカップ」がキーワード)、「ワールドカップ」の限定表現 s_{succ} は「日韓大会」となる。
- デ格 (s_{de})
キーワードがある用言の格要素であるとき、同じ用言を主辞とするデ格の格要素を限定表現として抽出する。例えば、「日韓大会で優勝した」という表現があったとき(「優勝」がキーワード)、「優勝」の限定表現 s_{de} は「日韓大会」となる。
- 近傍 (s_{dice})
近傍(同一文中)に存在する名詞のうちキーワードと関連が高い単語を限定表現候

補として全て取り出す。名詞は記事を茶釜を用いて形態素解析することで抽出する。キーワードと名詞間の関連度はコーパスにおける文書内の共起頻度に基づき、式 (4.1) の Dice 係数によって定義する。

$$D(x, y) = \frac{2|X \cap Y|}{|X| + |Y|} \quad (4.1)$$

- $|X|$ … 名詞 x が出現する記事数
- $|Y|$ … 名詞 y が出現する記事数
- $|X \cap Y|$ … 名詞 x, y が共に出現する記事数

以下に Dice 係数で限定表現候補を抽出する順序を述べる。

1. Dice 係数の計算

Dice 係数は毎日新聞の記事の 1991 年から 2003 年まで用いて計算した。そして、頻度が 100 以上でかつ $D(x, y) \geq 0.0005$ となるすべての x と y の組という条件下で $D(x, y)$ を事前に計算し、テーブルに格納した。以下、事前に計算した (x, y) の組ならびに Dice 係数の集合を C_{dice} とする。

2. キーワードと記事中の名詞についての Dice 係数が閾値以上ならば限定表現候補とする

キーワードと 3.2.2 項で抽出された記事中の名詞の組について C_{dice} から Dice 係数を求め、ある一定の閾値以上のときに限定表現候補として抽出する。閾値は 0.04 とした。しかし、複合名詞の場合は C_{dice} に単語の組が記載されていない可能性がある。したがって、単語が見つからなかった場合は単語の先頭 1 文字を削り、再び C_{dice} を検索する。これを単語が見つかるまで繰り返す。しかし、単語の長さを 1 文字まで削った場合にはたとえ Dice 係数が計算されている単語の組が見つかったとしても、その単語対が不適当な場合があるので、単語の長さは 2 文字以上とする。

● かぎ括弧 (s_{bkt})

かぎ括弧で囲まれた表現を抽出し、かつキーワードとの Dice 係数が 0.04 以上ならば限定表現とする。新聞記事では物の名前や作品のタイトルを示す固有表現は、かぎ括弧で表記されている。例えば、『スペースシャトル「ディスカバリー」』(スペースシャトルがキーワード) という表現があったとき「ディスカバリー」と「スペースシャトル」の Dice 係数が 0.13 なので、「スペースシャトル」の限定表現 s_{bkt} は「ディスカバリー」となる。

● 記事の先頭 (s_{front})

新聞は図 4.1 の下線が引かれているように記事の先頭にその記事のジャンルが書かれていることが多いので、記事の先頭が名詞ならば限定表現として抽出する。名詞は

サッカーのワールドカップ（W杯）日韓大会は6月30日、ブラジルの94年米国大会以来、史上最多の5回目の優勝で幕を閉じた。日韓がともに決勝トーナメントに進み、韓国はアジア初のベスト4進出を成し遂げた。一方で、優勝候補に数えられたフランス、アルゼンチン、イタリアなどが早期に敗退し、強豪国といえども、楽には勝ち進めなくなっていることを示した大会でもあった。5月31日にソウルでのフランス・セネガル戦でスタートした日韓大会の、1カ月にわたる戦いの跡を振り返る。

図 4.1: 記事の先頭例

文書解析で得られた記事を茶筌を用いて形態素解析することで抽出する。図 4.1 の場合「サッカー」が限定表現として抽出される。

以上の6つのパタンを図 4.2 に適用すると表 4.1 のような限定表現が得られる。図 4.2 において、キーワードは「*」で囲まれた「ワールドカップ」、「優勝」の2つである。解答タイプは国名であり、解答は四角で囲まれた「オーストリア」である。限定表現を抽出した結果、「解答」、「キーワード」、「限定表現」の3つの組が複数生成される。

ノルディックスキーの*ワールドカップ*（W杯）複合個人最終戦は10日、当地のホルメンコーレンで個人スプリントが行われ、前日初の総合*優勝*を決めたフェリックス・ゴットワルト（オーストリア）が前半ジャンプの6位から逆転して、今季6勝目を挙げた。

図 4.2: 限定表現抽出の例

4.2 属性の付与

本研究では、解答群の各解答に対するキーワードの限定表現がある程度似たような表現を持っていなければ、それらはキーワードの意味の曖昧性を適切に表現していないと判断する。そこで4.1節で抽出された（「解答」、「キーワード」、「限定表現」）の集合から、キーワードが共通で、かつその限定表現が共通の属性を持つものを選別し、これを解答群とする。属性とは限定表現が持つ単語の特徴のことを意味し、以下の8つの種類がある。なお、「かぎ括弧で囲まれた表現」、「末尾N文字」、「数量表現+接尾語」の3つは松本らのシステムと同じである。

- 数量表現+接尾語 (N)

数量表現と接尾語で構成されている限定表現を属性とみなす。例えば、「1998年」という限定表現は、「< NUM > + 年」という属性を持つ。数量表現には、漢数字とアラビア数字があるが、例えば「百」と「100」は表層表現としては違うがその意味は同じである。そこで、漢数字はすべてアラビア数字に直し、両者を同一とみなす。例

表 4.1: 表 4.2 の記事からの限定表現抽出結果

解答候補	キーワード	限定表現	パタン
オーストリア	ワールドカップ	ノルディックスキー	s_{front}
		ノルディックスキー	s_{dice}
		W杯	s_{dice}
		前半ジャンプ	s_{dice}
		複合個人最終戦	s_{dice}
		当地	s_{dice}
	優勝	ノルディックスキー	s_{front}
		ノルディックスキー	s_{dice}
		W杯	s_{dice}
		前半ジャンプ	s_{dice}
		複合個人最終戦	s_{dice}
		当地	s_{dice}
		前日初	s_{no}
		総合	s_{prev}

を表 4.2 に示す. ただし, 人名に漢数字がある場合は, 数字の意味ではないので漢数字のままとした.

表 4.2: 表記のゆれの吸収の例 (数量表現 + 接尾語 (N))

表記のゆれ吸収前		表記のゆれ吸収後	
限定表現	解答	限定表現	解答
百キロ級	井上康生	100 キロ級	井上康生
100 キロ級	井上康生		

- 末尾 N 文字 ($E1, E2, E3$)
 限定表現の末尾 1,2,3 文字をそれぞれ属性とみなす. 例えば, 「アルペンスキー」は, 末尾 1 文字が「 $E1:-$ 」, 2 文字は「 $E2:キー$ 」, 3 文字は「 $E3:スキー$ 」という属性を持つ.
- かぎ括弧で囲まれた表現 (K)
 限定表現がかぎ括弧で括られている名詞を属性とみなす. 新聞記事では物の名前や作品のタイトルを示す固有名詞は, かぎ括弧によって表記されている. そのため, かぎ括弧で括られている限定表現は固有名詞を示していると判断する. 例えば, 『スペースシャトル「ディスカバリー」』, 『大河ドラマの「毛利元就」』とあった場合, 「ディスカバリー」, 「毛利元就」といった表現がこの属性を持つ.

- シソーラスによる意味クラス (T)

限定表現の意味クラスを属性とみなす. シソーラスには角川類語新辞典 [11] を使用した. 限定表現をシソーラス上で検索し, 意味クラスを得て属性とする. ただし, 単に限定表現をシソーラスにかけるだけでは意味クラスが引けないことも多い. そこで, 意味クラスが引けない場合は単語の先頭一文字を削り, 再びシソーラスで検索する. これを順次繰り返す. ただし, 末尾 1 文字のときにシソーラスを引くと不適切な意味クラスが得られる可能性が高いため, 限定表現が 1 文字の場合は意味クラスはないものと判断する. 例えば, 「サッカー」という単語は「898d」の属性を持つ. 角川類語新辞典において「898d」は「球技」という意味クラスを表す.

- シソーラスによる上位の意味クラス (T')

限定表現が「意味クラス」属性を持つ場合, さらに上位の意味クラスをたどり抽出する. それを「上位の意味クラス」属性とみなす. 例えば, 「サッカー」という単語は「898d:球技」という属性 (T) を持つが, 「898d」のさらに上位語である「898」をたどって「スポーツ」という意味クラスを「上位の意味クラス」属性 (T') とする.

- 限定表現の抽出パターン (S)

4.1 項で述べた限定表現を抽出するパターンも解答群をまとめる属性とする. ただし, 近傍 (s_{dice}) と記事の先頭 (s_{front}) は解答群をまとめる属性とはしない. 近傍や記事の先頭によって抽出された限定表現は互いに関連がないことが多いので, 解答群としてふさわしくないからである.

- 連体修飾 (s_{no})
助詞「の」を介して取り出されたことを属性とみなす.
- 直前の単語 (s_{prev})
キーワードの直前の単語として取り出されたことを属性とみなす.
- 直後の単語 (s_{succ})
キーワードの直後の単語として取り出されたことを属性とみなす.
- デ格 (s_{de})
デ格の格要素として取り出されたことを属性とみなす.

表 4.3 に表 4.1 で得られた限定表現の属性を示す. 表中の N は数量表現 + 接尾語, $E1$ は末尾 1 文字, $E2$ は末尾 2 文字, $E3$ は末尾 3 文字, K はかぎ括弧, T はシソーラスの属性, T' は上位語のシソーラスの属性, S は限定表現抽出パターンが連体修飾か直前か直後かデ格なのかを表す. 「 ϕ 」は該当する属性が存在しないことを表す.

4.3 解答群の作成

4.1 節で述べた「解答候補 (a_i)」, 「キーワード (k_j)」, 「限定表現 (s_k)」の集合と 4.2 節で述べた「属性 ($attr$)」をもとに, キーワードと属性が共通しているもので解答群をまとめる.

表 4.3: 各限定表現の属性

限定表現	N	E1	E2	E3	K	T	T'	S
ノルディックスキー	$\varnothing$	ー	キー	スキー	$\varnothing$	898h,967	898	$\varnothing$
W杯	$\varnothing$	杯	W杯	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
前半ジャンプ	$\varnothing$	プ	ンプ	ャンプ	$\varnothing$	306	$\varnothing$	$\varnothing$
複合個人最終戦	$\varnothing$	戦	終戦	最終戦	$\varnothing$	282a	282	$\varnothing$
当地	$\varnothing$	地	当地	$\varnothing$	$\varnothing$	703	$\varnothing$	$\varnothing$
前日初	$\varnothing$	初	日初	前日初	$\varnothing$	$\varnothing$	$\varnothing$	s_{no}
総合	$\varnothing$	合	総合	$\varnothing$	$\varnothing$	267b	$\varnothing$	s_{prev}

ここでは考えられるすべてのキーワードと属性の組み合わせについて解答群を生成する. 例として「ワールドカップで優勝した国はどこですか」という質問のときに生成される解答群の候補を挙げる. 図 4.3, 図 4.4, 図 4.5 のような解答群が生成される. $AG(k, attr) = (s_i, a_i)$ は解答群を形式的に表現したものである. k はその解答群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キーワードである. s_i はキーワード k の限定表現, a_i はキーワード k を持つ解答候補, $attr$ は s_i が共通して持つ属性である.

図 4.3 は, キーワード「ワールドカップ」が共通で, 末尾 3 文字が「ー女子」を共通属性として持つ限定表現を含む (a_i, k_j, s_k) をまとめた解答群である. この解答群は, 「サッカー女子」, 「アルペンスキー女子」, 「ボブスレー女子」といったスポーツの種目に曖昧性があることを表している.

$$AG(\text{ワールドカップ}, E3: \text{ー女子}) = \{ (\text{サッカー女子}, \text{米国}), (\text{アルペンスキー女子}, \text{クロアチア}), (\text{ボブスレー女子}, \text{ドイツ}) \}$$

図 4.3: $AG(\text{ワールドカップ}, E3: \text{ー女子})$ の解答群

図 4.4 は, キーワード「ワールドカップ」が共通で, 「数 + 回」を共通属性として持つ限定表現を含む (a_i, k_j, s_k) をまとめた解答群である. 「1 回」, 「2 回」のように開催回数に曖昧性があることを表している.

$$AG(\text{ワールドカップ}, N: \text{数} + \text{回}) = \{ (1 \text{ 回}, \text{ウルグアイ}), (2 \text{ 回}, \text{日本}), (2 \text{ 回}, \text{メキシコ}), (2 \text{ 回}, \text{ドイツ}), (3 \text{ 回}, \text{日本}), (4 \text{ 回}, \text{フランス}), (4 \text{ 回}, \text{豪州}), (4 \text{ 回}, \text{スコットランド}), (5 \text{ 回}, \text{フランス}), (5 \text{ 回}, \text{南アフリカ}), (5 \text{ 回}, \text{オーストラリア}), (17 \text{ 回}, \text{フランス}), (34 \text{ 回}, \text{日本}) \}$$

図 4.4: $AG(\text{ワールドカップ}, N: \text{数} + \text{回})$ の解答群

図 4.5 は, キーワード「優勝」が共通で, 角川類語新辞典「898f」の意味クラス「水泳」を持つ限定表現を含む (a_i, k_j, s_k) をまとめた解答群である. この解答群は, どのような種目 (男子高飛び込み, 飛び込み, シンクロナイズドスイミング) で「優勝」したのかという曖昧性を表している.

$$AG(\text{優勝}, T:898f) = \{ (\text{男子高飛び込み}, \text{中国}), (\text{飛び込み}, \text{ロシア}), (\text{シンクロナイズドスイミング}, \text{ロシア}) \}$$

図 4.5: (優勝, T:898f) の解答群

ただし, 以下の場合には解答群を生成しない.

- 解答群を構成する要素が 1 つのとき
解答群を構成する要素が 1 つしかないグループはすでに解答が 1 つに決まっているので, 曖昧性がない. よって, 曖昧性を表す解答群としてふさわしくないので生成しない. 例を図 4.6 に挙げる.

$$AG(\text{ワールドカップ}, E3:\text{サッカー}) = \{ (\text{サッカー}, \text{フランス}) \}$$

図 4.6: 解答群を構成する要素が 1 つの例

- 解答群を構成する限定表現が全て同じとき
解答群中の限定表現が全て同じ場合, 問い返し文を生成して, ユーザから返答されても解答を 1 つに絞れない. よって, そのような場合は解答群を生成しない. 図 4.7 に挙げる. 図 4.7 の解答群を根拠にユーザにワールドカップのスポーツの種類を問い返しても, 限定表現は全て「サッカー」なので, 3 つの解答候補から正しいものを決めることはできない.

$$AG(\text{ワールドカップ}, E3:\text{サッカー}) = \{ (\text{サッカー}, \text{フランス}), (\text{サッカー}, \text{ブラジル}), (\text{サッカー}, \text{ドイツ}) \}$$

図 4.7: 解答群を構成する限定表現が全て同じ例

表記のゆれの吸収

表記のゆれの吸収とは, 表記が異なるが同じ意味を持つとき, その表記をまとめ 1 つの表現とすることである.

- 限定表現に表記のゆれがある場合

限定表現の表記が異なるが意味は同じであるとき、表記をまとめる必要がある。そこで、同じ解答が複数の限定表現を持つ場合、限定表現同士を後から 1 文字ずつ比較して、まったく同じ文字列が含まれていたら文字列の長いほうの限定表現を採用する。表 4.4 に表記のゆれの吸収の例を示す。解答「田村亮子」が 2 つの限定表現「48 キロ級」と「女子 48 キロ級」を持っている。2 つの限定表現を比較すると、「48 キロ級」が共通であり、文字列は「女子 48 キロ級」のほうが長いので、解答「田村亮子」が持つ限定表現は「女子 48 キロ級」のみとなる。

表 4.4: 限定表現の表記のゆれの吸収の例

吸収前		吸収後	
限定表現	解答	限定表現	解答
48 キロ級	田村亮子	女子 48 キロ級	田村亮子
女子 48 キロ級	田村亮子		

- 解答に表記のゆれがある場合

解答の表記が異なるが意味は同じであるとき、表記をまとめる必要がある。その処理は、解答タイプが「人名」であるか、そうでないかに分かれる。

- － 解答タイプが「人名」の場合

まったく同じ文字列が含まれていたら文字列の長いほうの解答を採用する。表 4.5 に例を示す。解答は「田村亮子」、「田村」、「亮子」である。限定表現はすべて同じ「女子 48 キロ級」なので、表記のゆれを吸収する。「田村」、「亮子」とも「田村亮子」に含まれる文字列なので、吸収され「田村亮子」となる。

表 4.5: 解答の表記のゆれの吸収の例 (解答タイプが人名)

吸収前		吸収後	
限定表現	解答	限定表現	解答
女子 48 キロ級	田村亮子	女子 48 キロ級	田村亮子
女子 48 キロ級	田村		
女子 48 キロ級	亮子		

- － 解答タイプが「人名」以外の場合

解答の後から 1 文字ずつ比較し、まったく同じ文字列が含まれていたら文字列の長いほうの解答を採用する。表 4.6 に例を示す。限定表現「第 1 ステージ」に対して、解答は「鹿島アントラーズ」、「アントラーズ」である。限定表現が同じなので、表記のゆれを吸収する。「アントラーズ」は「鹿島アントラーズ」の末尾の文字列が一致しているので、「鹿島アントラーズ」となる。

表 4.6: 解答の表記のゆれの吸収の例 (解答タイプが人名以外)

吸収前		吸収後	
限定表現	解答	限定表現	解答
第1ステージ	鹿島アントラーズ	第1ステージ	鹿島アントラーズ
第1ステージ	アントラーズ		

解答タイプが「人名」のときとそれ以外のときとでの処理の違いは, 解答タイプが人名のときは, 2つの解答が包含関係にあれば表記の統一を行うのに対し, 解答タイプが人名以外のときは, 2つの解答の末尾の文字列が一致しているときのみ表記の統一を行う点である.

4.4 解答群に対するスコア付け

一般に解答群は複数得られる. 問い返し文の内容を1つ決定するために, 複数の解答群にスコア付けする. 解答群に対するスコアを式(4.2)のように6つのサブスコア G, A, Z, AS, D, K の重み付き和で定義する.

$$Score(AG(k_j, attr)) = w_1 G + w_2 A + w_3 Z + w_4 AZ + w_5 D + w_6 K \quad (4.2)$$

$$G = \frac{G_{type}}{|GG|}, A = \frac{A_{type}}{|AG|}, AZ = \frac{\sum_{a \in AG} az(a)}{|AG|}, D = \frac{\sum_{key, gen \in AG} dice(key, gen)}{|GG| \times |key|}, K = \frac{\sum_{a \in AG} k(a)}{|GG|} \quad (4.3)$$

- G_{type} ... 解答群中の限定表現の異なり数
- $|GG|$... 解答群中の限定表現の数
- A_{type} ... 解答群中の解答の異なり数
- $|AG|$... 解答群中の解答の数
- $\sum_{a \in AG} az(a)$... 解答群中の解答のスコアの総和
- $az(a)$... 解答群中の解答 a に対して質問応答システムが与えるスコア
- $\sum_{key, gen \in AG} dice(key, gen)$... 解答群中のキーワードと限定表現のダイス係数の総和
- $dice(key, gen)$... 解答群中のキーワードと限定表現のダイス係数
- $|key|$... キーワードの数
- $\sum_{a \in AG} k(a)$... 解答群中の限定表現の抽出パタンのスコアの総数
- $w_1, w_2, w_3, w_4, w_5, w_6$... それぞれのスコアの重みで, それぞれ 0.05, 0.05, 0.2, 0.2, 0.2, 0.3 に設定

限定表現の異なり数についてのスコア (G)

分母は解答群中の限定表現数, 分子は解答群中の限定表現の異なり数である. もし, 限定表現の異なり数が少ない場合, 同じ限定表現候補が異なる解答を持つことが多いということの意味する. しかし, 適切に限定表現が抽出されているのであれば, 1つの限定表現に対して得られる解答は1つのはずである. この項は, 同じ限定表現が異なる解答に対して出現する場合に低いスコアを与える.

解答の異なり数についてのスコア (A)

分母は解答群中の解答数, 分子は解答群中の解答の異なり数である. もし, 解答の異なり数が少ない場合, 同じ解答が異なる限定表現を持つことが多いということの意味する. しかし, 適切に解答が抽出されているのであれば, 1つの解答に対して得られる限定表現は1つのはずである. この項は, 同じ解答が異なる限定表現に対して出現する場合に低いスコアを与える.

属性のスコア (Z)

限定表現の共通属性 *attr* に応じて与えられるスコアである. 表 4.7 に属性に応じて与えられるスコアの値を示す. スコアが高い属性ほど限定表現間の共通性が高く, 解答群として適切であるとみなしている.

表 4.7: 属性のスコア

属性		スコア
数量表現 + 接尾語		1
末尾3文字		0.8
末尾2文字		0.5
末尾1文字		0.2
意味クラス		0.7
上位の意味クラス		0.6
かぎ括弧		0.7
限定表現の抽出ボタン	連体修飾 (s_{no})	0.4
	直前 (s_{prev})	0.4
	直後 (s_{succ})	0.4
	デ格 (s_{de})	0.2

解答群中に存在する解答の平均スコア (AZ)

分母は解答群中の解答候補数, 分子は解答候補が持つスコアの総和である. この項は, 解答がどのくらい信頼性があるかを表しており, 高いスコアを持つ解答が解答群中にたくさんあればあるほど解答群としてスコアが高くなる.

キーワードと限定表現の Dice 係数の平均スコア (D)

キーワードと限定表現の Dice 係数の平均をスコアとしている. この項は, 限定表現が質問文中のキーワードとどのくらい関連しているかを表しており, 関連が高ければ高いほどスコアが高くなる.

限定表現の抽出パタンの平均スコア

分母は解答群中の限定表現の数, 分子は限定表現の抽出パタンのスコアの総和である. 表 4.8 に限定表現の抽出パターンごとに与えられるスコアを示す. この項は, 抽出された限定表現がどのくらい信頼できるかを表しており, 信頼できる抽出パターンで抽出される限定表現が多いほどスコアが高くなる.

表 4.8: 限定表現の抽出パタンのスコア

限定表現の抽出パターン	スコア
連体修飾 (s_{no})	1
直前 (s_{prev})	1
直後 (s_{succ})	1
デ格 (s_{de})	0.8
Dice 係数 (s_{dice})	0.5
記事の先頭 (s_{front})	0.5

4.5 本手法と松本らの手法との比較

本節では, 解答群作成の手法について, 松本らの方法に対する改善点をまとめる.

限定表現候補の抽出について

本手法では松本らの限定表現候補抽出に加えて, 記事の先頭の名詞を限定表現候補として抽出した. 新聞は記事の先頭にその記事のジャンルが書かれていることが多いからである.

属性の付与について

本手法では松本らの限定表現の属性として新たに、「上位語の意味クラス」、「連体修飾」、「直前の単語」、「直後の単語」、「デ格」を加えた。本研究では共通属性を持つ解答が1つのグループにまとめられるが、松本らの手法では限定表現が取り出されているにもかかわらず解答群としてまとまらないことがあった。例えば「サリン事件が起こったのはいつですか」という質問がある。質問のキーワードは「サリン事件」、「起きた」である。解答群としてふさわしいのは表 4.9 のような解答群である。

表 4.9: 「サリン事件が起こったのはいつですか」についての解答群

限定表現	解答
松本	1994年
地下鉄	1995年

新聞では、「松本サリン事件」、「地下鉄サリン事件」のような表現が現れる。限定表現の「松本」や「地下鉄」とはキーワードの「サリン事件」の前に現れる名詞である。松本らは「松本」と「地下鉄」をまとめるための属性を用意していなかったが、「直前」という属性を導入することで表 4.9 のような解答群を作成できるようになった。

表記のゆれの吸収

表記が異なるが意味は同じであるとき、表記をまとめる。具体的には以下の3種類の処理を行った。

- アラビア数字と漢数字の表記のゆれ

属性の「数量表現 + 接尾語」において、漢数字とアラビア数字では表層表現としては違いがあるが意味しているところは同じである。そこで、漢数字はすべてアラビア数字に直した。なお、人名の場合は漢数字のままにしている。

- 限定表現の表記のゆれ

限定表現の表記が異なるが意味は同じであるとき、表記をまとめる。同じ解答を持つ限定表現の末尾の文字列が一致しているときのみ表記の統一を行う。

- 解答の表記のゆれ

解答の表記が異なるが意味は同じであるとき、表記をまとめる。解答タイプが人名のときは、同じ限定表現を持つ解答が包含関係にあれば表記の統一を行う。解答タイプが人名以外のときは、同じ限定表現を持つ解答の末尾の文字列が一致しているときのみ表記の統一を行う。

解答群に対するスコア付けについて

松本らの解答群に対するスコア付けは式 (4.4) であった.

$$Score(AG(k_j, attr)) = w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4 \quad (4.4)$$

$$S_1 = \frac{|AG|}{A}, S_2 = \frac{A_{type}}{|AG|}, S_4 = \frac{\sum_{a \in AG} s(a)}{|AG|}$$

- A … 解答候補の総数
- $|AG|$ … 解答群が持つ解答候補の数
- A_{type} … 解答群中の解答の異なり数
- $\sum_{a \in AG} s(a)$ … 解答群に存在する解答のスコアの総和
- $s(a)$ … 解答候補 a に対して質問応答システムが与えるスコア
- w_1, w_2, w_3, w_4 … それぞれのスコアの重みで, それぞれ 0.3, 0.4, 0.2, 0.1 に設定

S_2, S_3, S_4 は本手法 (式 (4.2)) のそれぞれ G, Z, AZ と同じであり, 新しく取り入れたスコアは, A, D, K である. S_1 は取り入れなかった. 本手法では, 解答が限定表現と 1 対 1 に対応しているか (A), 解答群中の限定表現が質問文中のキーワードとどれだけ意味的に関連しているか (D), 信頼度の高い抽出パターンによって抽出された限定表現がどれだけ含まれているか (K), といった観点からそれぞれスコア A, D, K を導入した. 一方, S_1 は, 解答の数が多いグループに高いスコアを与えている. 解答の数が多ければよい解答群という考えは必ずしも妥当ではないと考えたからである.

第5章 問い返し文の生成

5章では解答群を作成した. 解答群は問い返し文の内容を決定する処理にあたる. この章では, 解答群から問い返し文の表層表現を生成する方法を述べる.

5.1 処理の流れ

問い返し文は, 基本的にテンプレートを用いて生成する. テンプレートを表 5.1, 5.2, 5.3 に示す. テンプレート A, B, C は, 解答群の限定表現の異なり数や「属性」によって使い分けられる. テンプレートの総数は 23 である. 表 5.1, 5.2, 5.3 における「適用条件」はそのテンプレートを使って問い返し文を生成するための条件である. 詳しくは 5.3 節で述べる.

- テンプレート A

テンプレート A は, 限定表現が 3 種類以上のとき, 解答群が持つ曖昧なキーワード k と問い返し主題 $wh + s$ を用いて生成するテンプレートである. 問い返し主題 ($wh + s$) とは, 解答群中の限定表現全体を表す名詞 (s) に「何」「何の」「どんな」といった疑問詞 (wh) をつけた表現のことである. 問い返し主題について詳しくは 5.2 節で述べる.

表 5.1: 問い返し文テンプレート A

適用条件			テンプレート
属性 (attr)	順序	サ変	
-	$s < k$	-	$wh + s$ の k ですか
数量表現 + 接尾語	$s < k$	-	$wh + s$ k ですか
末尾 N 文字	$s < k$	○	$wh + s$ で k したのですか
意味クラス	$k < s$	-	k の $wh + s$ ですか
上位語の意味クラス	$k < s$	○	k したのは $wh + s$ ですか

- テンプレート B

テンプレート B は, 限定表現が 3 種類以上のとき, 解答群が持つ曖昧なキーワード k のみを用いて生成するテンプレートである. 解答群によって問い返し主題が決まらないことがある. 例えば, 「スペースシャトルに搭乗した日本人はだれですか」という

質問の場合,「スペースシャトル」には「コロンビア」,「エンデバー」,「ディスカバリー」などのスペースシャトルの名前がある点で曖昧である.このとき「コロンビア」,「エンデバー」,「ディスカバリー」のいずれかを聞き出すような問い返し文を生成する.しかし,これらをまとめて参照する適切な表現がなく,先ほどのような問い返し主題を生成できない.そこで,このような場合はキーワードのみで問い返し文を生成する.さきほどの例は,図 5.2 のテンプレート B から「スペースシャトルの名前は何ですか」や「スペースシャトルの種類は何ですか」といった問い返し文が生成される.

表 5.2: 問い返し文テンプレート B

適用条件			テンプレート
属性 (attr)	順序	サ変	
かぎ括弧	-	-	k の名前は何ですか
	-	-	k の種類は何ですか
	-	○	k した名前は何ですか
	-	○	k した種類は何ですか
連体修飾	-	-	何の k ですか
	-	○	何の k したのですか
直前	-	-	何 k ですか
	-	○	何 k したのですか
直後	-	-	k の何ですか
	-	○	k したのは何ですか
デ格	-	-	何で k ですか
	-	○	何で k したのですか

● テンプレート C

テンプレート C は, 限定表現が 2 種類するとき, 二者択一の問い返し文を生成するテンプレートである. s_1, s_2 は 2 つの限定表現を表す. なお, 表 5.3 の限定表現 s_1 または s_2 の末尾に序数詞を含む場合, 限定表現の前に「第」をつけ, s_1 または s_2 として問い返し文テンプレート C に適用する. 序数詞のリストはあらかじめ人手で設定した. 序数詞の一覧を図 5.1 に示す. また, テンプレート C による生成例を図 5.2 に示す.

イニング, ケタ, ゲーム, コマ, シーン, セクション, セット, 回, 学年, 球, 局, 号, 次, 週, 打, 代

図 5.1: 序数詞一覧

図 5.3 に問い返し文生成の流れを示す. テンプレートによって問い返し文の候補を複数生成する. 問い返し文生成の流れは 2 つある. 限定表現候補が 2 種類の場合はテンプレート

表 5.3: 問い返し文テンプレート C

適用条件			テンプレート
属性 (attr)	順序	サ変	
-	s < k	-	s_1 の k ですか、 s_2 の k ですか
	s < k	-	s_1 k ですか、 s_2 k ですか
	s < k	○	s_1 で k したのですか、 s_2 で k したのですか
	k < s	-	k の s_1 ですか、 s_2 ですか
	k < s	-	k s_1 ですか、 s_2 ですか
	k < s	○	k したのは s_1 ですか、 s_2 ですか

質問文 = 芥川賞を受賞したのはだれですか

AG(芥川賞, N:数 + 回) = {(1 2 3 回, 松浦寿輝), (1 2 6 回, 長嶋有)}

問い返し文生成 = 第 1 2 3 回の芥川賞ですか、第 1 2 6 回の芥川賞ですか、
第 1 2 3 回芥川賞ですか、第 1 2 6 回芥川賞ですか

問い返し文選択 = 第 1 2 3 回の芥川賞ですか、第 1 2 6 回の芥川賞ですか

図 5.2: テンプレート C による生成例

C で生成し, 3 種類以上はテンプレート A または B で生成する. 生成テンプレートを適用する際には, 共通属性 attr, k と s の出現順序, k がサ変名詞であるかの 3 つの適用条件が考慮される. 詳しくは 5.3 節で述べる. また, テンプレート A を使うときは問い返し主題が必要である. 問い返し主題の生成方法については 5.2 節で述べる. テンプレートによって問い返し文を複数生成した後, n-gram の頻度に基づくスコア付けにより日本語として自然な文を選択する. スコアの詳細については 5.4 節で述べる.

5.2 問い返し主題の生成

問い返し主題とは, 「何」「何の」「どんな」といった疑問詞と解答群が持つ限定表現全体を指す表現を合わせたものである. 問い返し主題が生成されるのは, 解答群の限定表現が 3 種類以上であり, なおかつ, 限定表現全体を指す表現があるときのみである.

まず, 問い返し主題の候補をいくつか生成する. この処理は, 解答群の共通属性 attr によって以下のように分かれる. なお, 「1. 共通末尾の検出」の処理は共通属性にかかわらず全ての解答群に対して処理する.

1. 共通末尾の検出

- 解答群中の全ての限定表現候補を後から 1 文字ずつ伸ばしていき共通する部分を探す. その部分の前に疑問詞「何」「何の」「どんな」をつけ, 問い返し主題と

図 5.3: 問い返し文の生成の処理の流れ

する.

表 5.4 に「シドニー五輪のマラソンでメダルをもらったのはだれですか」という質問の例を示す. この場合は全ての限定表現に「メダリスト」が共通しているので, 「メダリスト」に対して疑問詞をつけて問い返し主題としたほうが適切である. そこで, 限定表現の共通末尾に対して疑問詞をつけ「何メダリスト」「何のメダリスト」「どんなメダリスト」といった問い返し主題の候補を生成する.

表 5.4: 共通末尾の例

キーワード:マラソン 属性:E3:リスト	
限定表現	解答
金メダリスト	高橋尚子
銀メダリスト	リディア・シモン
銅メダリスト	ジョイス・チェプチュンバ

- 解答群中の全ての限定表現候補を後から 1 文字ずつ伸ばしていき共通する部分文字列を探す. 共通しない部分が同じ意味クラスに属しているならば, 「< 疑問詞> + < 意味クラス名> の< 共通文字列> 」という問い返し主題の候補を作る. 表 5.5 に「シドニー五輪でメダルをもらったのはだれですか」という質問の例を示す. この場合は「金メダリスト」の前の部分「マラソン」「テニス」「砲丸投げ」が全て「スポーツ」の意味クラスに属している. そこで, 意味クラス名と共通末尾の前に疑問詞をつけ「何スポーツの金メダリスト」「何のスポーツの金メダリスト」「どんなスポーツの金メダリスト」といった問い返し主題の候補を生成する.

表 5.5: 意味クラス + 共通末尾の例

キーワード:シドニー五輪 属性:E3:リスト	
限定表現	解答
マラソン金メダリスト	高橋尚子
テニス金メダリスト	ヴィーナス・ウィリアムズ
砲丸投げ金メダリスト	アルシ・ハリユ

2. 共通属性が「数量表現 + 接尾語」のとき

接尾語の前に「何」という疑問詞をつけて問い返し主題とする. さらに, 接尾語が図 5.1 の序数詞のいずれかであるならば「第」をさらに前につけて問い返し主題とする. 例えば, 「芥川賞を受賞したのはだれですか」という質問のとき, 表 5.6 の解答群を得たとする. そのとき, 「回」の前に「何」という疑問詞をつけ, さらに「回」が序数詞なので「第」もつける. よって, 問い返し主題は「第何回」となる.

表 5.6: 解答群「芥川賞を受賞したのはだれですか」

キーワード：芥川賞 属性：N:数量表現 + 回	
限定表現	解答
1 2 3 回	松浦寿輝
1 2 6 回	長嶋有
1 2 7 回	吉田修一

また、接尾語の末尾の部分文字列が助数詞なら、その部分文字列に「何」をつけた表現を問い返し主題とする。末尾の部分文字列が助数詞であるかを判定するために、EDR 日本語単語辞書 [12] で品詞が「JN6(後置助数詞)」である 341 語と、分類語彙表 [13] で意味クラス「11962(助数接辞)」を持つ 412 語の和集合 563 語からなる助数詞のリストを作成した。さらに、部分文字列が図 5.2 の序数詞のいずれかならば「第」を前につけて問い返し主題の候補とする。

助数詞に「何」をつけて問い返し主題の候補とするのは、助数詞が接尾語の末尾の部分文字列である場合に限っている。この理由を以下に述べる。表 5.7 のような解答群があったとする。ここで、末尾に限らず接尾語の部分文字列に助数詞が含まれるならば、問い返し主題の候補を生成すると仮定する。まず、接尾語の部分文字列が助数詞のリストに存在するか調べる。助数詞のリストには「キロ級」は存在しないが「キロ」と「級」が存在するので、問い返し主題として「何キロ」と「何級」を生成する。「何級」は問い返し主題としてふさわしいが、「何キロ」はそうではない。これは、日本語の複合名詞の主辞が後ろにあるからである。よって、接尾語の末尾の部分文字列に対してのみ、助数詞に関する問い返し主題の候補を生成するようにした。

表 5.7: 接尾語の末尾の部分文字列を助数詞とする例

キーワード：柔道 属性：N:数量表現 + キロ級	
限定表現	解答
48 キロ級	田村亮子
60 キロ級	野村忠宏
100 キロ級	井上康生

3. 共通属性が「末尾 N 文字」のとき

末尾の前に「何」、「何の」、「どんな」という疑問詞をつけ、それらを問い返し主題とする。また、末尾の部分文字列が図 5.2 の序数詞のいずれかの場合は「第」を一番前につける。

例えば、「野球の新人王を受賞したのはだれですか」という質問のとき、表 5.8 の解答群を得たとする。そのとき、「末尾 3 文字」から「何」、「何の」、「どんな」という疑

問詞をつけ、「何リーグ」、「何のリーグ」、「どんなリーグ」という問い返し主題の候補を得る。

表 5.8: 解答群「野球の新人王を受賞したのはだれですか」

キーワード：新人王 属性：E3:リーグ	
限定表現	解答
セ・リーグ	上原浩治
パ・リーグ	松坂大輔
米大リーグ	佐々木主浩

ここで注意してほしいのは、共通属性が「末尾 N 文字」のとき、上記の処理による問い返し主題の候補と、1. 共通末尾の処理によって生成される問い返し主題の候補の両方がともに生成されるということである。例えば、表 5.5 の例では、その共通属性から「何のリスト」のような問い返し主題が、共通末尾を検出することで「何のメダリスト」のような問い返し主題の候補がともに生成される。この場合は後者の方が自然だが、解答群によっては前者のタイプの方が自然なときもある。

4. 共通属性が「意味クラス」、「上位語の意味クラス」のとき
意味クラスの前に「何」、「何の」、「どんな」という疑問詞をつけ、それらを問い返し主題とする。例えば、「世界選手権で優勝した国はどこですか」という質問のとき、表 5.9 の解答群を得たとする。そのとき、意味クラス「898」から「スポーツ」という意味クラス名を得て、「何」、「何の」、「どんな」という疑問詞をつけ、「何スポーツ」、「何のスポーツ」、「どんなスポーツ」という問い返し主題の候補を生成する。

表 5.9: 解答群「世界選手権で優勝した国はどこですか」

キーワード：世界選手権 属性：T:898d	
限定表現	解答
女子ソフトボール	米国
サッカー	フランス
バドミントン	インドネシア
マラソン	日本

問い返し主題の選択

複数生成された問い返し主題の候補に対し、Web の検索エンジンのヒット数を調べる。ヒット数が一番多いものを最終的に生成する問い返し主題とする。全てのヒット数が 0 に

なる場合は1つ決定することができないので、全てを問い返し主題とする。ヒット数の最も高いものが複数ある場合は、最も高いものの複数を問い返し主題とする。これらのときはそれぞれの問い返し主題に対してテンプレートが適用され、問い返し文の候補が生成される。なお、検索エンジンは「Google Web API」を使用した。

5.3 テンプレートの適用条件

テンプレート A,B,C には以下の3種類の適用条件があり、これを満たすときのみ適用される。

- 解答群の共通属性 (attr)

表 5.1, 5.2, 5.3 の「属性 (attr)」の列は、解答群中の限定表現の共通属性に関する適用条件を表す。例えば、表 5.1 には attr が「数量表現 + 接尾語」「末尾 N 文字」「意味クラス」「上位語の意味クラス」のときに適用されるテンプレートがある。

- キーワードと限定表現の出現順序

表 5.1, 5.2, 5.3 の「順序」の列は、解答群中のキーワードと限定表現の出現順序に関する適用条件を表す。「 $k < s$ 」はテンプレートのキーワード k が限定表現 s_i もしくは問い返し主題 $wh + s$ よりも前にあることを示す。「 $s < k$ 」は s_i もしくは $wh + s$ が k よりも前にあることを示す。テンプレートにおける k と s の順序関係は、解答 a_i を抽出した根拠文における k と s の順序関係と同じことが条件である。図 5.4 の根拠文から表 5.10 の解答群が得られたときを例に挙げる。図 5.4 において限定表現「オートバイ」「自転車」ともキーワード「世界選手権」と同一文中に現れ、さらに「 $s < k$ 」の順序になっている。このとき、適用条件が「 $s < k$ 」になっているテンプレート、すなわち限定表現がキーワードの前にあるような問い返し文を生成するテンプレートが適用される。

ここで問題となるのは、解答群の中でキーワードと限定表現の位置関係が解答候補毎に異なる場合である。そこで、式 (5.1) によって「根拠文における k と s の順序」を決定する。 $P > 0$ のとき、順序は「 $k < s$ 」となり、キーワードより問い返し主題が後になるテンプレートのみが適用される。 $P < 0$ のとき、順序は「 $s < k$ 」となり、キーワードより問い返し主題が前になる。 $P = 0$ のときは両方のテンプレートが適用される。式 (5.1) は、4.1 節の限定表現候補の抽出と同時に計算される。なお、同一文中にキーワードと限定表現候補がない場合は式 (5.1) の計算に使用しない。

$$P = -1 \cdot |f| + 1 \cdot |b| \quad (5.1)$$

- $|f|$ … 根拠文において限定表現がキーワードよりも前であるという解答候補の個数

オートバイの*世界選手権*シリーズ第15戦、オーストラリア・グランプリ（GP）は19日、フィリップアイランドで行われ、最高峰のモトGPクラスは既に年間総合優勝*を決めているバレンティーノ・ロッシ（イタリア、ホンダ）が41分53秒543（平均時速172・006キロ）で*優勝*した。

自転車*の*世界選手権*トラック種目第3日は1日、シュツットガルト（ドイツ）のベロドロームで行われ、男子スプリントで金子貴志（愛知）は準々決勝で敗れ、2日の5～8位決定戦に回った。

図 5.4: 表 5.10 の解答を含む根拠文

表 5.10: 解答群「世界選手権で優勝した国はどこですか」

キーワード：世界選手権 属性：T:997a	
限定表現	解答
オートバイ	イタリア
自転車	オーストリア

– $|b| \dots$ 根拠文において限定表現がキーワードよりも後であるという解答候補の個数

- 曖昧なキーワード k がサ変名詞

表 5.1, 5.2, 5.3 において、「サ変」の列が「○」のテンプレートは k がサ変名詞のときのみ適用される。 k がサ変名詞であるか否かは、 k の茶釜の辞書における品詞が「名詞-サ変接続」かどうかで判定する。また、複合名詞の場合には、 k がサ変名詞であっても「サ変」の欄が「○」になっているテンプレートを使用して問い返し文を生成しない。例えば、 k が「代表監督」の場合に茶釜の辞書における品詞は「代表」が「名詞-サ変接続」、「監督」も「名詞-サ変接続」である。しかし、「代表監督する」という表現はしない。このように複合名詞はサ変名詞によって構成されていても全体としてはサ変名詞とはならないことが多いので、「サ変」の欄が「○」になっているテンプレートを使用しない。また、 k がサ変以外なら「サ変」の欄が「○」になっているテンプレートと「-」になっているテンプレートの両方を使い、問い返し文を複数生成する。

5.4 n-gram の頻度によるスコアリング

表 5.1, 5.2, 5.3 のテンプレートにキーワード k と問い返し主題 $wh + s$ または限定表現 s_i をあてはめ、問い返し文を複数生成する。次に、n-gram の頻度によるスコア付けを行う。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する. ここで, n-gram の n はどこまでとればよいのか, 単語単位と文字単位の n-gram のどちらがよいのかという 2 つの選択肢がある. 文字 n-gram を用いたときのスコアを式 (5.2) に, 単語 n-gram を用いたときのスコアを式 (5.3) に示す. $S(q_{sys})$ は, 問い返し文の候補 q_{sys} に含まれる n-gram の頻度 $o(c_1^n), o(w_1^n)$ の重み付き和である. m-gram から 1-gram まで考慮し, n が大きいほど高い重みをつけた. このスコアは, 問い返し文中に含まれる部分文字列 (文字 n-gram または単語 n-gram) のコーパスにおける出現頻度が高ければ高いほど, その問い返し文は自然な表現であるという考えに基づいて定義されている.

$$S(q_{sys}) = \frac{1}{|q_{sys}|} \sum_{n=1}^m \left(w_n \cdot \sum_{c_1^n \in q_{sys}} O(c_1^n) \right) \quad (5.2)$$

$$S(q_{sys}) = \frac{1}{|q_{sys}|} \sum_{n=1}^m \left(w_n \cdot \sum_{w_1^n \in q_{sys}} O(w_1^n) \right) \quad (5.3)$$

- q_{sys} … 問い返し文の候補
- $o(c_1^n)$ … 文字 n-gram c_1^n の出現頻度 (新聞記事)
- $o(w_1^n)$ … 単語 n-gram w_1^n の出現頻度 (新聞記事)
- w_n … n-gram の n に対する重み

スコアを求める際に考慮する n-gram の最大長 m, ならびに n-gram を文字単位にするか単語単位にするかを定めるため, 以下のような実験を行った. まず, 表 5.13 に示す曖昧な質問を 28 個用意した. これらの質問に対して適切な解答群ができたと仮定し, その解答群を入力として問い返し文を生成した. その際, m を 1 から 7 まで変動させ, 最適な m を選んだ. また, 文字単位の n-gram (式 (5.2)) と単語単位の n-gram (式 (5.3)) の両方のスコアも比較した. 各 n-gram の頻度に対する重みは表 5.11 のように人手で設定した. 結果を表 5.12 に示す. 表中の数字は, スコアが 1 位となった問い返し文が適切であった事例数である. 適切な問い返し文とは, ユーザの質問の曖昧性を表現しており, なおかつ, 日本語として自然なものとした. m を 6 とし, かつ文字の n-gram を用いたときが 28 問中 26 問正解と最もよかった. よって, 本手法ではこのスコアを採用する. スコアの定義を改めて書くと式 (5.4) となる.

$$S(q_{sys}) = \frac{1}{|q_{sys}|} \sum_{n=1}^6 \left(w_n \cdot \sum_{c_1^n \in q_{sys}} O(c_1^n) \right) \quad (5.4)$$

$$w_6 = 0.9, w_5 = 0.09, w_4 = 0.009, w_3 = 0.0009, w_2 = 0.00009, w_1 = 0.00001$$

表 5.11: 予備実験での n-gram の頻度に対する重み

m	w_1	w_2	w_3	w_4	w_5	w_6	w_7
1	1	-	-	-	-	-	-
2	0.1	0.9	-	-	-	-	-
3	0.01	0.09	0.9	-	-	-	-
4	0.001	0.009	0.09	0.9	-	-	-
5	0.0001	0.0009	0.009	0.09	0.9	-	-
6	0.00001	0.00009	0.0009	0.009	0.09	0.9	-
7	0.000001	0.000009	0.00009	0.0009	0.009	0.09	0.9

表 5.12: 予備実験の結果 (28 問中)

m	1	2	3	4	5	6	7
単語 n-gram	23	22	23	23	23	24	24
文字 n-gram	23	23	24	23	24	26	24

5.5 自然な問い返し文を生成するための工夫

この節では自然な問い返し文を生成するために工夫した点をまとめる。

- n-gram の頻度によるスコア付け

日本語として自然な文を選択するために、新聞記事での文字の n-gram の頻度を用いる。新聞には大量の自然な日本語が載っており、問い返し文に含まれる n-gram の頻度が高い方がより自然な日本語であるとみなす。例えば、図 5.5 のような複数の問い返し文の候補が生成されたとき、n-gram に基づくスコアが高い (2) が選ばれる。

- (1) 地下鉄のサリン事件ですか、松本のサリン事件ですか
 (2) 地下鉄サリン事件ですか、松本サリン事件ですか

図 5.5: どちらが日本語として正しい文か？

- Web の検索エンジンのヒット数の利用

基本的には n-gram の頻度によって文の自然さを評価するが、この手法は疑問詞の選択については有効ではない。新聞には、「何」、「何の」、「どんな」という疑問詞はほとんど現れないからである。そのため、疑問詞を含む問い返し主題を生成するときは Web の検索エンジンのヒット数を調べ、最もヒット数の高いものを日本語として自然な問い返し主題とした。

表 5.13: 予備実験の質問一覧

質問	質問文	質問の曖昧な情報
1	ボクシングの世界チャンピオンはだれですか	階級, 日付
2	サッカーの日本代表はだれですか	性別, 何歳以下, 日付
3	クラブユースサッカー選手権で優勝したのは	何歳以下, 日付
4	百メートルの世界記録保持者はだれですか	性別, スケート, 陸上, 水泳
5	セリエAの日本人はだれですか	サッカー, バレー, 日付
6	サリン事件がおきたのはいつですか	地下鉄, 松本
7	日本シリーズで優勝したのはだれですか	野球, ゴルフ, 将棋, 日付
8	西カンファレンスで優勝したのはだれですか	NBA, NHL, 日付
9	棋聖戦で優勝したのはだれですか	囲碁, 将棋, 日付
10	全英オープンで優勝したのはだれですか	テニス, ゴルフ, 日付
11	全米オープンで優勝したのはだれですか	テニス, ゴルフ, 日付
12	ワールドカップで優勝したのはだれですか	サッカー, ラグビー, スキー, 日付
13	NHK杯で優勝したのはだれですか	フィギアスケート, 体操, 将棋, 日付
14	日本グランプリで優勝したのはだれですか	F1, オートバイ, 日付
15	水泳の世界選手権で優勝したのはだれですか	種目, 日付
16	グランドスラムを達成したのはだれですか	テニス, ゴルフ, 囲碁, 日付
17	アカデミー賞を受賞したのはだれですか	部門, 日付
18	グラミー賞を受賞したのはだれですか	部門, 日付
19	レコード大賞を受賞したのはだれですか	部門, 日付
20	芥川賞を受賞したのはだれですか	日付
21	直木賞を受賞したのはだれですか	日付
22	芸術選奨文部大臣賞を受賞したのはだれですか	部門, 日付
23	六大学野球で優勝した大学はどこですか	東京, 関西, 春季, 秋季, 日付
24	野球の新人王は誰ですか	セリーグ, パリーグ, 大リーグ, 日付
25	タイガースの監督は誰ですか	阪神, デトロイト, 日付
26	高校野球で優勝したのはどこですか	春, 夏, 日付
27	甲子園で優勝したのはどこですか	春, 夏, 阪神, 日付
28	世界選手権で優勝した国はどこですか	スポーツ, 日付

- キーワード k と限定表現 s_i または問い返し主題 $wh + s$ の位置

図 5.14 のようなキーワード k と問い返し主題 $wh + s$ の位置だけが入れ替わっている 2 つの問い返し文から 1 つを選択するために、文字 n-gram の頻度を用いると、「世界チャンピオンの何級ですか」が選択される。しかし、どちらかといえば「何級の世界チャンピオンですか」の方が自然である。これらの文は k と $wh + s$ の位置が入れ替わっているだけなので、文字 n-gram の頻度によるスコアではほとんど差がつかない。そこで、 k と $wh + s$ の位置関係を決めるために、解答候補を抽出した根拠文における両者の並びを参照した。すなわち、問い返し文におけるキーワードと限定表現（または問い返し主題）の順序と根拠文における両者の順序が同じ方が問い返し文として自然であると考えた。そのため、両者の順序が一致しないときには問い返し文を生成しないというテンプレートの適用条件を設けた。

表 5.14: 「世界チャンピオン」 k と「何級」 $wh + s$ の位置の違い

問い返し文	式 (5.4) のスコア
何級の世界チャンピオンですか	539
世界チャンピオンの何級ですか	584

第6章 評価実験

6.1 実験方法

3,4,5章で述べた問い返し文を生成する手法を評価するために、曖昧な質問を50問用意し問い返し文を生成できるか実験を行った。曖昧な質問の用意の仕方を以下に述べる。人手で曖昧性のある質問を最初に用意し、次にキーワードで記事を検索し、複数の解答候補が抽出されるかどうか判断した。解答が抽出された場合は曖昧な質問として実験に使用した。本研究で取り扱う解答群生成から問い返し文生成までの処理を評価し、またその際、他のモジュールの誤りの影響を排除するため、解答が抽出されない場合は評価用の質問とはしなかった。質問の一覧を付録Aの表A.1,A.2,A.3,A.4に示す。

質問解析は自動で行っていないため、文書検索、解答抽出に必要なプライマリキーワード、セカンダリキーワード、検索年、解答タイプ、キーワードタイプを付録Aの表A.5,A.6に示すように人手で設定した。検索年の「all」とは1991年から2003年までの13年の文書の検索を意味している。

4章の手法によって式(4.2)が最大となる解答群を選択した。そこで得られた解答群が適切であるかどうかを人手で調べた。ここで解答群が適切であるということは以下の2つを指す。

- 得られた解答群の限定表現がキーワードの意味の曖昧性を誤りを含まず正しく反映すること
- 得られた解答群が正しい解答を含むこと

表6.1にどのくらい適切な解答群を抽出できたかについて本手法と松本らの手法との比較実験した結果を示す。実験条件は表A.1,A.2,A.3,A.4の同じ質問50問を与えた。表6.1の(a)は曖昧性を解消するための解答群が獲得できなかった数である。(b)は適切な解答群が最上位のスコアを持つ数である。(c)は適切な解答群が獲得できたがスコアが最上位でない数である。(d)は適切な解答群が獲得できたときに、その解答群が平均してどの順序にあるかを表す。

表6.2に5章の手法によって問い返し文を生成した実験結果を示す。なお、問い返し文生成の精度を評価するために、表6.1の(b)と(c)の適切な解答群46個が全て最上位になったと仮定した。正解としてふさわしい問い返し文は、ユーザの質問の曖昧性を表現しており、なおかつ日本語として自然なものとした。表6.2の「検索」は複数の問い返し主題をWebの検索エンジンのヒット数で1つに選択するとき、適切でない問い返し主題のヒット数

表 6.1: 解答群作成の実験結果

	本手法	松本らの手法
適切な解答群が存在しない (a)	4(8%)	15(30%)
適切な解答群が得られるかつ解答群が最上位 (b)	32(64%)	17(34%)
適切な解答群が得られるかつ解答群が最上位以外 (c)	14(28%)	18(36%)
適切な解答群が得られたときの平均順位 (d)	4.2 位	15 位

の方が大きく, 適切な問い返し主題が選ばれなかったことを指す. 「位置」はキーワードと限定表現または問い返し主題の位置が適切でないことを指す. 「n-gram」は複数の問い返し文から n-gram の頻度に基づくスコアで問い返し文を選択するモジュールで間違えたことを指す. 「他」は検索, 位置, n-gram 以外の原因で誤り, 適切な問い返し文が得られなかったことを指す.

表 6.2: 問い返し文の生成の実験結果

正解	不正解			
35	11			
	検索	位置	n-gram	他
	1	1	6	3

3,4,5 章の手法によって, 文書検索から問い返し文生成まで連続して処理したときの実験結果を表 6.3 に示す. 表 6.3 の「2 位以下」とは適切な解答群があるが最上位のスコアを持たないことを指す. 「なし」とは適切な解答群が得られなかったことを指す.

表 6.3: 本手法の実験結果

解答群が適切 かつ 問い返し文が適切	解答群が不適切または問い返し文が不適切					
	23					
	解答群が不適切		問い返し文が不適切			
27	18		5			
	2 位以下	なし	検索	位置	n-gram	他
	14	4	0	1	4	0

6.2 考察

6.2.1 解答群作成の考察

表 6.1 の結果から, 松本らの手法より本研究の手法の方がよいことがわかる. 以下に例を挙げながら表 6.1 の結果を分析し考察を述べる.

解答群抽出成功例

「ボブスレー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図 6.1 の適切な解答群が抽出された. ボブスレーのワールドカップは「男子」と「女子」の区別があるので適切な解答群と判断した. 図 6.2 に図 6.1 が抽出された記事を示す. キーワードは「*」で囲まれた「ボブスレー」, 「優勝」の 2 つ, 解答は四角で囲まれた「ドイツ」, 「スイス」である. 図 6.2 に解答「ドイツ」, 限定表現「女子」の記事が 3 つあるが, 1 つにまとまり解答群に表示される. 解答群は 4.2 節で述べた属性が上位の意味クラス「511:男女」でまとまっている. 男子が持つ意味クラスは「511a:男」であり, 女子が持つ意味クラスは「511b:女, 512:幼児, 513a:少女」である. これは松本らの手法では, 同じ意味クラスがないので解答群としてまとまらなかった. 本手法では, 上位の意味クラスも共通属性として用いるので, また女子も男子も上位の意味クラス「511:男女」を持つので, 図 6.1 のように解答群としてまとまった. 詳細の「G」「A」「Z」「AS」「D」「K」は 4.4 節の解答群に対するスコア付けによるサブスコアである. これらを式 (4.2) に代入したときの解答群のスコアが 0.5569 である. これは図 6.1 の 1 行目にも示される. また, 図 6.1 の「1 位」はこの解答群の順位である.

———1 位のスコア キーワード:ボブスレー 属性:T:511 score:0.5569———
———詳細:G:0.67 A:0.67 Z:0.60 AS:0.30 D:0.0531 K:1.00———
女子:ドイツ
男子:ドイツ
男子:スイス

図 6.1: 適切な解答群の抽出例

適切な解答群は獲得できているが, 最上位でない例

適切な解答群が獲得できているが, 最上位にならなかった解答群の例を図 6.3 に示す. 「東カンファレンスで優勝したチームはどこですか」という質問に対して, 東カンファレンスには「米プロバスケットボール」と「北米プロアイスホッケー」があるので, 12 位の解答群が適切だと判断した. 図 6.4 に図 6.3 が抽出された記事を示す. 図 6.4 の下線部の限定表

*ボブスレー*女子の*ワールドカップ*(W杯)開幕戦最終日は11日、ドイツのウィンターベルクで行われ、日本B(羽栗尚美、吉村美鈴)は合計1分59秒75で、1位から5秒21遅れで27チーム中の21位、日本A(檜野真奈美、山本晴美)は1分59秒86で22位だった。ザンドラ・プロコフがパイロットのドイツAが、初戦に続いて*優勝*。

*ボブスレー*女子の*ワールドカップ*(W杯)第2戦最終日は18日、ドイツのケーニッヒスゼーで行われ、日本B(桧野、吉村)が合計タイム1分43秒55で、1位から3秒60遅れで参加25チーム中の19位、日本A(羽栗、山本)は1分45秒39で23位だった。*優勝*はドイツB。

*ボブスレー*女子の*ワールドカップ*(W杯)第6戦は23日、オーストリアのイグルスで行われ、日本A(桧野、山本)は合計タイム1分55秒30で22位、日本B(羽栗、吉村)は1分58秒21で26位だった。*優勝*はドイツAで1分51秒51。

男子*ボブスレー*の*ワールドカップ*(W杯)は17日、イタリアのコルティナダンペッツォで4人乗りを行い、ドイツAが合計タイム1分45秒60で*優勝*。日本(鈴木、石井、清水、大滝)は25位だった。

*ボブスレー*男子の*ワールドカップ*(W杯)第7戦は25日、フランスのラプレーニュで2人乗りを行い、ソルトレークシティー五輪代表の日本A(鈴木寛、井上将憲)は2回の合計2分1秒33で21位、日本B(大石博暁、土井川真二)は同2分2秒76で27位だった。スイスAが同1分59秒06で今季初*優勝*した。

図 6.2: 図 6.1 の記事

現「米プロバスケットボール」はキーワード「東カンファレンス」との Dice 係数が高いので限定表現として抽出される。一方、「北米プロアイスホッケー」は本手法の 4.1 節の限定表現抽出「記事の先頭」から抽出された限定表現である。これは松本らの手法では抽出されない。なぜなら、キーワードの「東カンファレンス」「優勝」に対して限定表現抽出パターン「連体修飾」「直前」「直後」「デ格」のいずれにも当てはまらず、「近傍」での抽出パターンを用いても、表 6.4 に示すように事前に計算された名詞の組とその Dice 係数の集合 C_{dice} (4.1 節) に登録されていないため、やはり抽出されない。すなわち、本手法の「記事の先頭」の抽出パターンの有効性が確かめられた。一方、解答群をまとめる属性は、上位の意味クラス「898:スポーツ」である。しかし、スコアが低いために 1 位になっていない。原因はサブスコアの「K:限定表現の抽出パターンのスコア」が極端に低いからである。これは「記事の先頭」から取り出された限定表現は信頼度が低いとされているからである。

———1 位のスコア 優勝 E2:ぶり score:0.5434———
———詳細:G:1.00 A:1.00 Z:0.50 AS:0.19 D:0.0281 K:1.00———
4 季ぶり:スパーズ
1 2 年ぶり:レーカーズ
4 年ぶり:ネッツ
———12 位のスコア 東カンファレンス T':898 score:0.3962———
———詳細:G:0.50 A:1.00 Z:0.60 AS:0.23 D:0.0268 K:0.50———
北米プロアイスホッケー:マイティダックス
バスケットボール:セルティックス
米プロバスケットボール:レーカーズ
米プロバスケットボール:ネッツ
米プロバスケットボール:スパーズ
北米プロアイスホッケー:デビルズ

図 6.3: パタン「記事の先頭」によって抽出された限定表現を含む解答群の例

表 6.4: 限定表現とキーワードの Dice 係数

	東カンファレンス	カンファレンス	優勝
北米プロアイスホッケー	なし	なし	なし
プロアイスホッケー	なし	なし	なし
アイスホッケー	なし	0.011	0.004

米プロバスケットボール、NBAのプレーオフは2日、カリフォルニア州サクラメントで西カンファレンス決勝（7回戦制）の第7戦が行われ、レーカーズが延長の末、112—106でキングズを振り切り、3年連続のファイナル進出を決めた。レーカーズは3連覇をかけ、*東カンファレンス*優勝*の**ネッツ**と戦う。北米プロアイスホッケー、NHLの年度王者を決めるスタンリー・カップ決勝（7回戦制）は9日、当地のペプシセンターで最終戦を行い、西カンファレンス覇者の地元アバランチが3—1で*東カンファレンス*覇者の**デビルズ**を破り、5シーズンぶり2度目の*優勝*を飾った。

図 6.4: 図 6.3 の記事

解答群抽出失敗例

50 問中 4 問の質問が適切な解答群を作ることができなかった。その要因を以下に示す。

「自由民主党の党首はだれですか」という質問に対して、図 6.5 の解答群が得られた。自由民主党には、英国やソ連の他に日本がある。英国やソ連の自由民主党では「党首」だが、日本の自由民主党の「党首」のことを「総裁」という。「党首」も「総裁」も同じ意味であるが、表記が違うので文書検索に失敗したことが原因である。さらに、新聞では「自由民主党」よりも「自民党」と書くほうが多いので「自由民主党」というキーワードよりも「自民党」というキーワードのほうがより多くの記事が検索できる。このように、1 つの意味を表す表現が複数あると文書検索の際に問題となる。

———19 位のスコア 自由民主党 T:709c score:0.5111———
 ———詳細:G:0.67 A:1.00 Z:0.70 AS:0.43 D:0.0045 K:0.67———
 英国:アシュダウン
 ソ連:ジリノフスキー
 英国:ブレア

図 6.5: 日本が入っていない解答群

そこで、仮に「党首」と「総裁」や「自由民主党」と「自民党」の表記のゆれを解決したと仮定して、次の実験を行った。「自由民主党の総裁はだれですか」「自民党の党首はだれですか」「自民党の総裁はだれですか」という 3 つの質問に対して、限定表現「日本」が含まれている解答群を得ることができれば、「自由民主党の党首はだれですか」の解答群とあわせれば適切な解答群が得られることになる。しかし、限定表現「日本」を得ることができなかった。新聞で「自由民主党」や「自民党」と記載されていれば、日本の自由民主党を指すことが当たり前になっているからである。

「共産党の党首はだれですか」という質問も「自由民主党の党首はだれですか」と同じことが原因で適切な解答群を得ることができなかった。共産党の党首のことを「幹部会委員長」や「総書記」と表現することや、新聞で「共産党」と記載されていれば日本の共産党を指すことから、「日本」という限定表現を抽出できなかった。

「スノーボードのワールドカップで優勝したのはだれですか」という質問に対して、適切な解答群が得られなかったのは、図 6.5 に示すように「スノーボード」や「ワールドカップ」といったキーワードとスノーボードの種目との間の Dice 係数が事前に用意した C_{dice} の中にないか、存在しても Dice 係数が 0.04 より低く限定表現として抽出されないことにある。「大回転」と「スノーボード」の Dice 係数は 0.04 より高いが、「大回転」だけ抽出されても解答群をつくることはできない。

表 6.5: スノーボードの種目の Dice 係数

	スノーボード	ワールドカップ	優勝
ハーフパイプ	なし	なし	なし
スラローム	0.018	0.003	なし
大回転	0.092	0.029	0.010

「野手部門の最優秀選手はだれですか」という質問も「スノーボードのワールドカップで優勝したのはだれですか」と同じことが原因で適切な解答群を得ることができなかった。「野手部門」「最優秀選手」というキーワードに対して、「セ」「パ」「大リーグ」が C_{dice} に存在しないことや存在しても Dice 係数が 0.04 より低く限定表現として抽出されなかった。

以上のことから以下の 4 つのことが起こると適切な解答群を得ることができないことがわかった。

- 表記のゆれに対応していない
- 適切な限定表現が記事中に存在しない
- キーワードと限定表現の間の Dice 係数を求めることができない
- 両者の Dice 係数が 0.04 より低く限定表現として抽出されない

6.2.2 問い返し文の生成の考察

表 6.2 の結果について、以下に例を挙げながら結果の分析と考察を述べる。

問い合わせ文の生成の成功例

「ソルトレークシティー五輪スケートの金メダリストはだれですか」に対して、図 6.6 の適切な問い合わせ文が生成された。「スケート」には「スピード」と「フィギュア」の曖昧性がある。それが、意味クラス「898h:スキー」でまとめられ、限定表現が 2 種類なので、テンプレート C によって「スピードのスケートですか、フィギュアのスケートですか」と「スピードスケートですか、フィギュアスケートですか」が生成される。式 (5.4) のスコアによって「スピードスケートですか、フィギュアスケートですか」が選択された。

———1 位のスコア スケート T:898h score:0.5734———
———詳細:G:0.40 A:1.00 Z:0.70 AS:0.38 D:0.0847 K:0.90———
スピード:ドーン・ルメイ・カトリオナ
フィギュア:マリナ・アニシナ
フィギュア:アレクセイ・ヤグディン
スピード:クラウディア・ペヒシュタイン
スピード:西谷岳文
———問い合わせ文候補———
スピードのスケートですか、フィギュアのスケートですか
スピードスケートですか、フィギュアスケートですか
———問い合わせ文選択———
スピードスケートですか、フィギュアスケートですか

図 6.6: 問い合わせ文の成功例

問い合わせ文の生成の失敗例 (表 6.2 の「検索」)

「ゴールデングラブ賞を受賞したのはだれですか」という質問に対して図 6.7 の解答群が得られた。「ゴールデングラブ賞」には投手部門や三塁手部門など様々な部門がある。限定表現の種類が 3 種類以上なので、表 6.8 の問い合わせ主題を作成し、検索エンジンにより 1 つ選択する。最も自然な問い合わせ主題は「何部門」であるが、ヒット数が最も多い「どんな手部門」が選択された。最後にテンプレート A によって「ゴールデングラブ賞のどんな手部門ですか」が問い合わせ文として生成された。

——9位のスコア ゴールデングラブ賞 E2:部門 score:0.3555——

——詳細:G:0.50 A:0.67 Z:0.50 AS:0.16 D:0.0711 K:0.50——

セ・リーグ外野手部門:松井
投手部門:松井
三塁手部門:大村
セ・リーグ外野手部門:松坂
投手部門:松坂
三塁手部門:立浪

——問い合わせ文候補——

ゴールドデングラブ賞のどんな手部門ですか

——問い合わせ文選択——

ゴールドデングラブ賞のどんな手部門ですか

図 6.7: Web 検索のヒット数の失敗例 (解答群)

表 6.6: Web 検索のヒット数の失敗例 (Web 検索のヒット数)

問い合わせ主題	ヒット数
どんな手部門	1240
何手部門	0
何の手部門	0
どんな部門	378
何部門	578
何の部門	37
どんな門	50
何門	157
何の門	26

問い合わせ文の生成の失敗例 (表 6.2 の「n-gram」)

「松井が所属した球団はどこですか」という質問に対して、図 6.7 の問い合わせ文が生成される。次に式 (5.4) のスコアによって「松井の稼頭央内野手ですか、秀喜外野手ですか」、「松井稼頭央内野手ですか、秀喜外野手ですか」の 2 つから「松井の稼頭央内野手ですか、秀喜外野手ですか」が選択された。しかし、この場合はスコアが低い「松井稼頭央内野手ですか、秀喜外野手ですか」の方が自然である。

表 6.7: 問い合わせ文の生成の失敗例

問い合わせ文	n-gram の頻度によるスコア
松井の稼頭央内野手ですか、秀喜外野手ですか	30.9
松井稼頭央内野手ですか、秀喜外野手ですか	28.9

問い合わせ文の生成の失敗例 (表 6.2 の「位置」)

「ボブスレー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、図 6.8 の上位の意味クラス「511:男女」を共通属性とする解答群が得られた。ボブスレーのワールドカップには男女の区別があるので適切な解答群であると判断した。しかし、キーワードに対して限定表現の位置を決定するモジュールにおいて、解答群中の解答を含む根拠文においては限定表現はキーワードの後にあると判断され、テンプレート C によって図 6.9 の四角で囲まれた問い合わせ文「ボブスレーの女子ですか、男子ですか」のみが生成される。しかし、キーワードと限定表現の位置が逆の「女子のボブスレーですか、男子のボブスレーですか」が自然な日本語である。

———1 位のスコア ボブスレー T:511 score:0.5569———
———詳細:G:0.67 A:0.67 Z:0.60 AS:0.30 D:0.0531 K:1.00———
(限定表現:解答:位置)
女子:ドイツ:後
男子:ドイツ:前
男子:スイス:後

図 6.8: キーワードと限定表現の位置の失敗例 (解答群)

問い合わせ文の生成の失敗例 (表 6.2 の「他」)

「カー・オブ・ザ・イヤーを受賞したのは」という質問に対して、図 6.10 の上位の意味クラス「709:世界」でまとまった解答群が得られた。「カー・オブ・ザ・イヤー」には「欧

ボブスレーの女子ですか、男子ですか
ボブスレー女子ですか、男子ですか

図 6.9: キーワードと限定表現の位置の失敗例 (問い合わせ文生成, N-gram 選択)

州」「日本」「北米」があるので適切な解答群であると判断した。限定表現が3種類なので、表 6.10 の問い合わせ主題を生成する。しかし、どの問い合わせ主題を選択しても自然な日本語とはいえない。最後に図 6.11 の四角で囲まれた「どんな世界のカー・オブ・ザ・イヤーですか」がユーザに提示されるが、やはり自然な日本語とはいえない。「欧州」「日本」「北米」をまとめた表現で「どこの国のカー・オブ・ザ・イヤーですか」が自然な日本語である。

この問題を解決するために、場所のときに「どこ」を用いた問い合わせ文を生成するテンプレートを加えることが考えられる。さらに、限定表現が人の場合は「だれ」、時間を表現する場合「いつ」を用いた問い合わせ文を生成するテンプレートに加えるとより自然な問い合わせ文が生成できる可能性が高い。

——— 2 位のスコア カー・オブ・ザ・イヤー T':709 score:0.5347 ———
——— 詳細:G:0.67 A:1.00 Z:0.60 AS:0.20 D:0.0069 K:0.97 ———
欧州:日産
0 1 年日本:三菱自動車工業
日本:富士重工業
0 1 年日本:本田技研
日本:トヨタ
北米:フォード

図 6.10: 他の失敗例 (解答群)

表 6.8: 他の失敗例 (Web 検索のヒット数)

問い合わせ対象表現	ヒット数
どんな世界	22800
何世界	131
何の世界	660

どんな世界のカー・オブ・ザ・イヤーですか どんな世界カー・オブ・ザ・イヤーですか

図 6.11: 他の失敗例 (問い返し文選択)

6.2.3 解答群の選択処理と問い返し文の生成処理の融合

表 6.3 から解答群作成の正解率が悪いために問い返し文生成全体の正解率が下がっていることがわかる。現在は、解答群を選択した後、問い返し文を生成するという処理の流れである。しかし、解答群の候補と問い返し文の候補を同時に生成し、自然な問い返し文が得られれば、その解答群に対して高いスコアを与えることで解答群作成の正解率の向上が期待できる。そこで、適切な解答群が生成した問い返し文の n-gram のスコアと 1 位の解答群が生成した問い返し文の n-gram のスコアの関係性を調べた。しかし、n-gram のスコアの値が高くても解答群として不適切な場合や、逆にスコアが低くても適切な場合があり、現在の問い返し文生成のためのスコアを解答群のスコアに反映させるのは難しいことがわかった。

第7章 結論

本論文では、ユーザからの曖昧な質問が入力されたとき、解答候補から曖昧性を検出し、曖昧な情報をユーザに問い返す対話型質問応答システムについて述べた。同システムにおける要素技術のうち、主に解答群の作成と問い返し文の生成を行った。本研究では以下に述べるような手法で松本らによる解答群作成の手法を改善した。

解答群の作成

- 記事のジャンルが述べられている先頭を抽出
文書検索によって得られた記事の先頭には、その記事のジャンルを指す名詞が多くあることが多い。そこで記事を茶釜を用いて形態素解析し、先頭が名詞ならば限定表現候補とみなし取り出した。これによって、「セリアA」についての記事がほとんど「サッカー」について書かれていて「バレーボール」のように記事が少なく Dice 係数が低いために取り出せなかった名詞が限定表現として抽出されるようになった。
- 属性の追加
限定表現の抽出パターンを考慮した属性「連体修飾」「直前」「直後」「デ格」、さらに「意味クラス」の上をたどる属性「上位の意味クラス」を追加した。
- スコア付け
解答群中の限定表現や解答候補の異なり数、属性、解答の平均スコア、キーワードと限定表現との関連度、限定表現の抽出パターンという6つの観点からスコア付けを行った。

本手法と松本らの手法と比較実験を行った結果、適切な解答群が最上位のスコアを持つ事例が50問中32問となり、松本らの手法よりも大幅により成果を得ることができた。

次に、以下に述べる手法で問い返し文を生成した。

問い返し文生成

解答群中の限定表現の種類によって、問い返し文の生成が異なる。限定表現が3種類以上の場合、問い返し主題を生成し、キーワードと問い返し主題の位置を決定し、テンプレートによって問い返し文を生成する。限定表現が2種類の場合、キーワードと限定表現の位置を決定し、テンプレートによって問い返し文を生成する。

- 問い返し主題生成
解答群中の限定表現の共通する部分と疑問詞を組み合わせた表現を問い返し主題として生成する。
- キーワードと限定表現または問い返し主題の位置の決定
根拠文のキーワードと限定表現の位置を利用して、問い返し文における両者の位置を決定する。解答群中の限定表現は複数あり位置情報も前と後の両方ある可能性があるため、多数決によって決定する。同数の場合は位置を決定しない。
- テンプレート
解答群中の限定表現によって、用意した 23 種類のテンプレートからいくつかを選択し、問い返し文を生成する。
- n-gram の頻度による選択
n-gram の頻度によって、複数の問い返し文から 1 つを選択する。

以上の手法より問い返し文生成の実験を行い、46 問中 39 問 (81%) 正解した。文書検索から問い返し文生成まで全体を通して実験を行った結果は 50 問中 28 問 (56%) が正解だった。

今後の課題

正解率をさらに上げるために以下のことを提案する。

- 解答候補抽出の改善
文書検索し、解答候補を抽出するときに、解答を間違えて取り出す場合や適切な解答を取り出せない場合がある。文書が過去形で書かれている事柄から解答を抽出することや否定を表す言葉が表 3.5 のパタンで取り出そうとしている文章中にあるかを考慮することを提案する。
- 解答群中の限定表現の選定
本手法は解答群にスコアをつけることによって、適切な解答群は上位になるようにした。しかし、ほとんど適切であるといってもよい解答群でもふさわしくない限定表現が含まれることで不適切になる場合が多い。そこで解答群の限定表現を選定し削除することを提案する。例えば、「シドニー五輪の柔道の金メダリストはだれですか」という質問に対して表 7.1 の解答群が得られたとする。表 7.1 の下から 2 行目に「アトランタ五輪 71 キロ級」という不適切な限定表現が含まれている。「アトランタ五輪 71 キロ級」は「文字列 + 数量表現 + 末尾」という形をしている。解答群中の他の限定表現は「数量表現 + 末尾」の形をしている。そこで、どちらの形が多いか多数決をとり、少ないほうの形をした限定表現は削除する。すると、表 7.1 の「選定、削除後」の適切な解答群が得られる。

表 7.1: 解答群中の限定表現の選定

選定, 削除前		選定, 削除後	
限定表現	解答	限定表現	解答
48 キロ級	田村亮子	48 キロ級	田村亮子
60 キロ級	野村忠宏	60 キロ級	野村忠宏
100 キロ級	井上康生	100 キロ級	井上康生
アトランタ五輪 71 キロ級	中村兼三		

- 問い返し文生成テンプレート
6.2.2 項で述べたように, 人を問う質問には「だれ」, 場所を問う質問には「どこ」, 時間を問う質問には「いつ」をテンプレートに追加することを提案する.
- 解答群の選択方法の改良
6.2.3 項で述べたように, 生成された問い返し文の良さを, 解答群のスコアに反映させることができれば, 解答群生成の正解率の向上が期待できる. 現在の問い返し文の表層表現の良さを評価する n-gram の頻度によるスコアでは適切な解答群の選択には貢献しないことがわかったが, 他の手法を検討する余地はあると考えられる.

謝辞

本研究を進めるにあたり、白井清昭 助教授ならびに島津明 教授には、数多くの御教示を頂きました。また、本研究に関して多大な助言をしていただいた中村誠 助手に心から感謝致します。そして、島津・白井研究室の皆様方には、研究に関する貴重な支援をして頂きましたことを心より感謝致します。

参考文献

- [1] 松本匡史. 質問の曖昧性を考慮した質問応答システムに関する研究. 北陸先端科学技術大学院大学, 情報科学研究科, 情報処理学専攻, 修士論文, 2006.3
- [2] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi IBM's Statistical Question Answering System In 9th Text REtrieval Conference, Gaithersburg, MD, 2000.
- [3] 石塚満, 倉田岳人, 岡崎直観. 大域知能へ向けての係り受け関係に基づく グラフ構造を用いた質問応答機構
- [4] 清田陽司, 黒橋禎夫, 木戸冬子. 大規模テキスト知識ベースに基づく自動質問応答 - ダイアログナビ -. 2003
- [5] 黒橋禎夫, 日笠亘, 藤井綱貴 入力質問と知識ベースとの柔軟なマッチングに基づく対話的ヘルプシステム 2001 年情報学シンポジウム
- [6] Matthias Denecke, Hajime Tsukada. Instance-Based Generation for Interactive Restricted Domain Question Answering Systems. IJCNLP 2005 486-497
- [7] 徳江英範. 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. 北陸先端科学技術大学, 情報科学研究科, 情報処理学専攻, 修士論文, 2005.3.
- [8] IREX 実行委員会 (編). IREX ワークショップ予稿集. IREX 実行委員会, 1999.
- [9] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸. 日本語形態素解析システム「茶筌」 version 2.2.1 使用説明書 <http://chasen.aist-nara.ac.jp/chasen/bib.html.ja>, Dec. 2000.
- [10] Taku Kudo, Yuji Matsumoto Fast Methods for Kernel-Based Text Analysis, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, July 2003, pp. 24-31.
- [11] 大野 晋, 浜西 正人. 「角川類語新辞典」角川書店, 1981.
- [12] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. 1995. TR-045.
- [13] 国立国語研究所. 分類語彙表. 大日本図書. 2004

付 録 A 実験に用いた質問一覧

表 A.1: オープンテストの質問一覧 1

質問 番号	質問文
	質問の曖昧な情報
1	万博はどこで開催されましたか
	日付
2	府中市の市長はだれですか
	東京都・広島県, 日付
3	N H K 紅白歌合戦の司会はだれですか
	白組・赤組, 日付
4	ディズニーランドが開園したのはいつですか
	東京・ユーロ・香港
5	共産党の党首はだれですか
	日本・中国・ソ連, 日付
6	自由民主党の党首はだれですか
	日本・英国・ソ連, 日付
7	箱根駅伝で優勝したチームはどこですか
	日付
8	全日本学生スキー選手権で優勝したのはだれですか
	種目, 男女, 日付
9	世界水泳選手権において平泳ぎで優勝したのはだれですか
	距離, 男女, 日付
10	シドニー五輪において走り幅跳びで優勝したのはだれですか
	男女
11	全日本フィギュアスケート選手権で優勝したのはだれですか
	種目, 男女, 日付
12	全米オープンシングルスで優勝したのはだれですか
	男女, 日付

表 A.2: オープンテストの質問一覧 2

質問 番号	質問文
	質問の曖昧な情報
13	バドミントンの世界選手権でシングルスで優勝したのは
	男女, 日付
14	野球の首位打者はだれですか
	セ・パ・大リーグ・社会人, 日付
15	野球の盗塁王はだれですか
	セ・パ・大リーグ・社会人, 日付
16	サッカーの得点王はだれですか
	高校・Jリーグ, 五輪・選手権, 日付
17	シドニー五輪で体操の個人総合で優勝したのはだれですか
	男女
18	全日本選手権で新体操の個人総合で優勝したのはだれですか
	男女, 日付
19	レスリングフリースタイルで優勝したのはだれですか
	階級, 日付
20	皐月賞で優勝したのは
	日付
21	新人王戦で優勝したのは誰ですか
	囲碁・将棋, 日付
22	ファイナルファンタジーが発売されたのはいつですか
	7,8,9,X
23	東カンファレンスで優勝したチームはどこですか
	NBA・NHL, 日付
24	天皇杯を獲得したのはだれですか
	サッカー・テニス・レスリング, 日付
25	全国高校ラグビーフットボール大会で優勝したチームは
	日付
26	カー・オブ・ザ・イヤーを受賞したのは
	日本・北米・欧州, 日付
27	カンヌ国際映画祭でパルムドールを受賞したのは
	日付
28	ベネチア国際映画祭で金獅子賞を受賞したのは
	日付
29	ベルリン国際映画祭で金熊賞を受賞したのは
	日付

表 A.3: オープンテストの質問一覧 3

質問 番号	質問文
	質問の曖昧な情報
30	高校野球大会の入場行進曲を歌った歌手はだれですか
	日付
31	関東大学リーグで優勝したチームはどこですか
	サッカー・アメリカンフットボール, 日付
32	野球連盟の理事長はだれですか
	関西六大学・福岡六大学・高校野球, 日付
33	アジアカップで優勝した国はどこですか
	サッカー・卓球, 日付
34	松井が所属した球団はどこですか
	秀喜・稼頭央, 日付
35	新幹線が開業したのはいつですか
	東海道・山陽・東北, 日付
36	NECがVリーグで優勝したのはいつですか
	男女, 日付
37	フェンシングの団体で優勝した国はどこですか
	フルーレ・サーブル・エペ, 男女, 日付
38	アジア選手権の重量挙げで優勝したのはだれですか
	スナッチ・ジャーク, 階級, 男女, 日付
39	アーチェリーの世界選手権で優勝したのはだれですか
	男女, 日付
40	アトランタ五輪で飛び込みで優勝したのはだれですか
	飛び板飛び込み・高飛び込み, 高さ
41	カーリングのパシフィック選手権で優勝したのは
	男女, 日付
42	ボブスレーのワールドカップで優勝した国はどこですか
	2人乗り・4人乗り, 男女, 日付
43	砲丸投げの世界陸上で優勝したのはだれですか
	男女, 日付
44	ハンマー投げの世界陸上で優勝したのはだれですか
	男女, 日付
45	中田が年間最優秀選手に選ばれたのはいつですか
	英寿・浩二, 日付
46	野手部門の最優秀選手はだれですか
	セ・パ大リーグ, 日付

表 A.4: オープンテストの質問一覧 4

質問 番号	質問文 質問の曖昧な情報
47	スノーボードのワールドカップで優勝したのはだれですか ハーフパイプ・パラレル回転・スノーボードクロス, 日付
48	ソルトレークシティー五輪スケートの金メダリストは スピードスケート・フィギュアスケート, 日付
49	ゴールデングラブ賞を受賞したのはだれですか 投手部門・三塁手部門・外野部門, セ・パ・大リーグ, 日付
50	日本有線大賞を受賞したのは誰ですか 大賞・最多リクエスト歌手賞・最多リクエスト曲賞・最優秀新人賞, 日付

表 A.5: オープンテストの質問設定 1

質問 番号	primary keyword	secondary keyword	検索年	answer type	keyword type
1	万博	開催	2000-2003	loc	hyponym
2	府中市	市長	2000-2003	per	hyponym
3	司会	N H K 紅白歌合戦	99-2003	per	hyponym
4	開園	ディズニーランド	all	time	agent
5	党首	共産党	91	per	hyponym
6	党首	自由民主党	all	per	hyponym
7	優勝	東京箱根間往復大学駅伝競走	2000-2003	org	hyponym
8	優勝	全日本学生スキー選手権	all	per	agent
9	優勝	世界水泳選手権, 平泳ぎ	all	per	agent
10	優勝	シドニー五輪, 走り幅跳び	2000	per	agent
11	優勝	全日本フィギュアスケート選手権	all	per	agent
12	優勝	全米オープン, シングルス	2000	per	agent
13	優勝	バドミントン, シングルス, 世界選手権	all	per	agent
14	首位打者	野球	all	per	hyponym
15	盗塁王	野球	all	per	hyponym
16	得点王	サッカー	all	per	hyponym
17	優勝	シドニー五輪, 体操, 個人総合	2000	per	agent
18	優勝	新体操, 個人総合, 全日本選手権	all	per	agent
19	優勝	レスリング, フリースタイル	2000	per	agent
20	優勝	皐月賞	2000-2003	per	agent

表 A.6: オープンテストの質問設定 2

質問 番号	primary keyword	secondary keyword	検索年	answer type	keyword type
21	優勝	新人王戦	2000-2003	per	agent
22	発売	ファイナルファンタジー	all	time	hyponym
23	優勝	東カンファレンス	2000-2003	per	agent
24	優勝	天皇杯	2000-2003	per	agent
25	優勝	全国高校ラグビーフットボール大会	2000-2003	org	agent
26	受賞	カー・オブ・ザ・イヤー	99,2000	org	agent
27	受賞	カンヌ国際映画祭, パルムドール	all	ot	hyponym
28	受賞	ベネチア国際映画祭, 金獅子賞	all	ot	hyponym
29	受賞	ベルリン国際映画祭, 金熊賞	all	ot	agent
30	入場行進曲	高校野球大会	2000-2003	per	hyponym
31	優勝	関東大学リーグ	all	org	agent
32	理事長	野球連盟	2000-2003	per	hyponym
33	優勝	アジアカップ	all	na	agent
34	所属	松井	2000-2003	org	agent
35	開業	新幹線	92	time	agent
36	優勝	Vリーグ, NEC	2000-2003	time	agent
37	優勝	フェンシング, 団体	all	na	agent
38	優勝	重量挙げ, アジア選手権	2000-2003	per	agent
39	優勝	アーチェリー, 世界選手権	all	per	agent
40	優勝	飛び込み, アトランタ五輪	all	per	agent
41	優勝	カーリング, パシフィック選手権	all	na	agent
42	優勝	ボブスレー, ワールドカップ	all	na	agent
43	優勝	砲丸投げ, 世界陸上選手権	all	per	agent
44	優勝	ハンマー投げ, 世界陸上選手権	all	per	agent
45	年間最優秀選手	中田	all	time	hyponym
46	野手部門	最優秀選手	all	per	hyponym
47	優勝	スノーボード, ワールドカップ	2000-2003	per	agent
48	金メダリスト	ソルトレークシティ, 五輪, スケート	2002	per	hyponym
49	受賞	ゴールデングラブ賞	2000-2003	per	agent
50	受賞	日本有線大賞	all	per	agent