

Title	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
Author(s)	松本, 匡史
Citation	
Issue Date	2006-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1986
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 修士

修 士 論 文

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

松本 匡史

2006 年 3 月

修 士 論 文

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

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

410111 松本 匡史

提出年月: 2006 年 2 月

概 要

本論文は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、まず質問文に含まれるキーワードについて、そのキーワードの意味を限定するような表現（限定表現）を抽出する。次に共通の属性を持つ限定表現と同時に出現する解答のグループを作り、得られた解答群を曖昧なキーワードの意味と共に表示する。曖昧な質問 31 個に対して解答群を生成する実験を行い、74 % 程度の質問に対して適切な解答群が生成されることを確認した。

目次

第1章	序論	1
1.1	研究の目的	1
1.2	研究の背景	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	リスト型質問応答システムに関する研究	4
2.3	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	8
3.1	質問応答システムの概要	8
3.2	処理の流れ	9
3.2.1	質問文解析	9
3.2.2	文書検索	11
3.2.3	解答候補抽出	12
第4章	解答リストの作成	16
4.1	限定表現候補の抽出	17
4.2	属性の付与	20
4.3	解答群の作成	21
4.4	解答群に対するスコア付け	24
4.5	解答の出力	27
第5章	評価実験	28
5.1	実験方法	28
5.2	考察	28
第6章	結論	42

目 次

3.1	提案システムの理想的な出力例	8
4.1	パターンマッチで限定表現を抽出できない例	17
4.2	限定表現候補の抽出例	18
4.3	現時点のシステムの出力例	27
4.4	システムの理想的な出力例	27
5.1	正しい解答群の抽出例 (ボクシング)	31
5.2	解答群が上位に現れない例 (ノーベル賞の 1 位)	32
5.3	正しい解答群が上位に現れない例 (ノーベル賞の 13 位)	33
5.4	正しい解答群が上位に現れない例 (柔道の 1 位のスコア)	34
5.5	正しい解答群が上位に現れない例 (柔道の 3 位のスコア)	35
5.6	正しい解答群が上位に現れない例 (レコード大賞の 1 位のスコア)	37
5.7	正しい解答群が上位に現れない例 (レコード大賞の 3 位のスコア)	37
5.8	正しい解答群が上位に現れない例 (水泳の 1 位のスコア)	38
5.9	正しい解答群が上位に現れない例 (水泳の 17 位のスコア)	39

表 目 次

3.1	解答タイプ	10
3.2	キーワードタイプ	11
3.3	解答候補が満たすべき条件	12
3.4	解答候補抽出パターンとスコア	13
3.5	品詞情報のスコア	14
4.1	(a_j, k_i, s_k) の 3 つ組の例	19
4.2	各限定表現の属性	21
4.3	AG(アカデミー賞, 数 + 回) の解答群	22
4.4	AG(アカデミー賞, 男優賞) の解答群	22
4.5	AG(アカデミー賞, 括弧) の解答群	23
4.6	AG(アカデミー賞, 1857) の解答群	23
4.7	解答が複数の限定表現にまたがる例	25
4.8	属性のスコア	26
4.9	限定表現が持つ解答とスコアの例	26
5.1	実験結果	28
5.2	質問一覧	29
5.3	質問設定	30
5.4	質問に対する正解の限定表現と抽出された限定表現のリスト	40

第1章 序論

1.1 研究の目的

質問応答システムでは、ユーザが入力した質問に対して適切な解答を返すことが課題となっている。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がある。質問応答システムの現在の主流は、TRECやQACなどの評価型ワークショップ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の一問一答型のシステムである。これに対し、本研究では、ユーザの質問が曖昧であるために複数の解答候補が正解となる場合を考慮した質問応答システムの開発を目指す。例えば「ワールドカップの優勝国はどこですか」という質問が挙げられる。これはワールドカップがどのスポーツの大会を指すかが曖昧であるため、解答を1つに絞ることはできない。そこで、キーワードの意味の曖昧性を検出し、個々のキーワードの意味ごとに解答を提示する。また、ユーザが自分の質問の曖昧性に気付いていないとき、一問一答型のシステムによって解答を1つだけ提示すると、解答が誤りであっても正解であると誤認識する可能性がある。これに対し本研究では、キーワードの意味に曖昧性があることをユーザに知らせることによりユーザの誤認識を防ぐ。

1.2 研究の背景

現在主流の一問一答型の質問応答システムでは、複数の解答候補が得られた場合、解答候補を1つだけ出力することが一般的である。しかし、以下の2つのケースのように、解答を1つに絞り込むことが必ずしも適切であるとは限らない。

ユーザは複数の解答を要求している

質問が曖昧で複数の解答が得られるが、ユーザが求める解はその中の1つである

の例として「世界3大珍味は何ですか」というものが挙げられる。これは「キャビア」「トリュフ」「フォアグラ」という3つの解答を完答して正解となる質問である。の例としては「アテネ五輪の柔道の金メダリストは誰ですか」というものがある。これは柔道の階級が曖昧なため解答を1つに絞ることができない質問である。

本研究では主に の場合を取り扱う。曖昧性を解消する手法の例として、ユーザに対して問い返しを行う方法があるが、自然な問い返し文を生成することは困難である。また

ユーザは必ずしも 1 つの解答を求めているとは限らない。これに対し本研究では，曖昧なキーワードの意味とそれに対する解答をリストとして提示する。これによりユーザは適切な解答を容易に見つけることができる。

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，質問応答システムの関連研究や本研究の特色について述べる。
- 3 章では，質問応答システムの全体の構成について簡潔に述べる。
- 4 章では，ユーザの質問の曖昧性を検出するための手法を述べる。
- 5 章では，提案手法の評価実験とその考察について述べる。
- 6 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

質問応答システムとは人間が普段使っている自然言語によって入力された質問に答えるシステムである。近年の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反面、情報が増大したことによって、正しい情報を入手するまでに時間がかかる場合が多い。そのような状況下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ので容易に情報を入手することができる。1990年代ころから、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。本節では、過去の質問応答システムに関する研究をいくつか概観し、本研究との違いについて述べる。

2.1 一般的な質問応答システム

現在の質問応答システムの基本的な流れは以下の通りである。

質問文の解析による、キーワードと解答タイプの抽出

キーワードによる文書検索

解答抽出

質問応答システムの現在の主流は、TREC や QAC などの評価型ワークショップに代表されるように、ユーザの質問に対して複数の解答が得られた場合でも、その中の1つが正しいければ正解であると判断し、システムが1つの解答を返すだけの一問一答型のシステムである。一問一答システムではシステム全体の精度を上げるために、上に挙げた処理を行うサブシステムの改良がいくつか試みられている。

Hidaka らは、従来の質問応答システムで質問解析時に用いられる5種類の解答タイプ (Person, Location or Organization, Date, Numeral, Others) では、解答候補がたくさん得られすぎて正しい解答を検出するのは困難であると考え、解答タイプを200種類に増やし正しい解答を抽出することを試みた [1]。解答タイプは質問の意図を知るための重要な情報である。例えば、解答タイプ「Person」は質問文の「誰ですか?」の意味を持つ。増やした解答タイプは階層構造になっており、階層を上にとどると最終的に従来の5種類の解答タイプにたどり着く。また、「Other」というのは他の4つのタイプに属さないタイプのことを指す。解答タイプを増やしたことで、例えば質問文の解析の結果が5種類の解答タイプでは「Others」に属していたものが、200種類の解答タイプを用いることで「Person」

のサブタイプに属することがある．これにより，従来は質問タイプが「other」と判定された質問の意図が「誰ですか」であることが分かる．

Deepak らは解答を抽出するために，ウェブから解答抽出のパターンを自動的に学習するシステムを構築した [2]．例えば「When was X born?」のような質問があった場合，解答として「Mozart was born in 1756.」が得られたなら，「< NAME > was born in < BIRTHDATE >」のように正規表現としてパターン化する．更に得られたパターンを 6 つの解答タイプ (BIRTHDATE, LOCATION, INVENTOR, DISCOVER, DEFINITION, WHY-FAMOUS) に分類した．すると，BIRTHDATE, LOCATION については正しい解答タイプを抽出することができた．しかし，DEFINITION タイプのパターンマッチを用いると，パターンに適合する表現が多すぎて正しい解答タイプを獲得することができなかったと報告している．

2.2 リスト型質問応答システムに関する研究

リスト型質問応答システムとは，質問に応じて解答候補を複数表示するシステムである．現在主流の一問一答型のシステムは，知識源となる文書集合の中から与えられた質問に対して解答候補群を見つけ，解答の信頼性であるスコアを付与し，上位のスコアの解答を出力する．従って解答が複数得られた場合においても，最上位の解答を正しい解答として出力する．例えば「日本三景は何ですか」という質問に対して従来のシステムを用いると「安芸の宮島」「天橋立」「松島」のいずれかを出力する．これらはいずれも日本三景の 1 つで正解である．しかしユーザの求める意図としては「三景」とあるように 3 つの解答すべてを知りたいことが伺えるため，システムが解答を 1 つしか出力しないのは問題である．そこで近年複数の解答をもつ質問にも対応する質問応答システムの研究も行われ始めた．

石下らは複数の回答を持つ質問を処理するために，解答候補の集合のスコア分布が正解集合に対するスコア分布と不正解集合のスコア分布の混合分布であると仮定し，それらの 2 つの分布を EM アルゴリズムによって分離して，正解側の分布に由来すると推定できる回答候補を正解として出力している [8]．また質問応答システムは一般的に不得意な質問が存在するが，そのような場合には得られた回答候補群のスコアを再度計算し，回答候補群を改めて順位付けしている．これによりリスト型質問応答の評価指標である F 値の精度を向上させることができたが，不正解の回答も正解集合に加えてしまうという傾向が見られた．

Fukumoto らは質問応答システムによって抽出されたトップ 20 の解答候補を用いて解答のリストを作成した [3]．解答候補の中に含まれている正解の解答，不正解の解答を識別するために，以下の方法でスコア付けしている．

- 同じ文の中にある解答候補のスコアのウェイト
- キーワードと一番近い位置にある解答候補のウェイト

- 上記の2つの方法によって計算されたスコアと質問応答システムで得られたスコアの和

更に解答の数を質問の表層表現で確認している．例えば「three animal」は解答数が3，「combi」は解答数が2，「who and who」は2であることを示す．

2.3 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解答を選択するものである．

Chaiらは，現実の世界では質問が単独で行われるよりも，関連のある質問を順序立てて行うことに着目し，連続して質問応答が行うことができるシステムを提案した [7]．例えば「古代都市ポンペイを破壊した火山は何ですか」という質問に対しての解答は「ヴェスビオ」であるが，もしユーザが「それはいつ起こりましたか」のように，一番初めの質問と関連のある質問を尋ねた場合「それ」が指すのは「破壊」であることをシステムが自動的に認識し，解答「74年」を出力するものが挙げられる．「それ」のような照応詞は，談話役割（質問中の名詞句や動詞句が，質問のトピックや焦点が何なのかを示す情報）と談話変化（談話役割がある質問から別の質問に話題が変わるときにどのように変化するかという情報）を用いて決められる．しかし，このシステムは実装までは至っていない．

清田らはパーソナルコンピュータの Windows 環境の利用者を対象とした対話的質問応答システムを構築した [9]．データベースにはマイクロソフトがすでに保有している膨大なテキスト知識ベースをそのままの形で利用している．ユーザが曖昧な質問をしたときに対話的に聞き返しを行うことでユーザの具体的な回答をナビゲートする．ただ，質問が曖昧すぎると回答が含まれる文を取り出すことが困難となるため，聞き返しの文の生成ができない．そこで清田らは「対話カード」というものを作成した．対話カードとはあらかじめ質問を用意しておいて，ユーザの質問が対話カードとマッチした場合，対話カードに従って聞き返しを行うものである．対話カードに存在しない質問に対しては，ユーザの質問と知識ベース中の文とマッチした複数の文からそれぞれ状況説明文（知識ベース中の文とマッチしなかった部分に，ユーザの問題を具体化する状況説明が書いてある部分）を抽出し，ユーザに提示することで曖昧性を解消している．例えば「WindowsXP 上でページ違反が発生する」という質問に対して，知識ベースの「WindowsXP で IE6 を起動した際にページ違反が発生する」や「WindowsXP でディスククリーンアップを実行するとページ違反が発生する」という文とマッチした場合「IE6 を起動した」「ディスククリーンアップを実行した」の部分を抽出する．ただ，このシステムは Windows に関するドメインに限定される．

徳江らはデータベースに毎日新聞の記事12年分を用いたドメインに依存しない対話的質問応答システムに関する研究を行った [6]．曖昧性の判断は，質問文を解析して得られたキーワードの意味を限定するような表現（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、そのキーワードは曖昧であるとみなしている。限定表現を抽出するために、あらかじめキーワードにかかる限定表現のパターン（連体修飾句、格助詞デ格でかかる名詞、キーワードが複合名詞を作る際の直前・直後の単語、日付表現）を記述しておき、そのパターンにマッチした限定表現を抽出する。例えば、「金メダリスト」がキーワードのとき、「柔道の金メダリスト」という文から、キーワードを連体修飾している「柔道」が限定表現として抽出される。そして、キーワードと抽出された限定表現の種類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の検出を行う。ここで限定表現が持つ特徴（属性：「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」）を利用して、より適切な限定表現のグループを作成している。例えば、「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、「数量表現＋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している。そして得られたいくつかの限定表現のグループをスコア付けし、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は数が少なく、問い返し文生成までには至っていない。

2.4 本研究の特色

今までに挙げた関連研究と本研究の違いは以下の通りである。

- ユーザの曖昧な質問に対応する

2.1 節で紹介した質問応答システムとの違いとして、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挙げられる。曖昧であるというのは、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。例えば、「アテネオリンピックの柔道の金メダリストは誰ですか」という質問が入力された場合、柔道には60キログラムや48キログラムといった階級が14個存在するため、解答が一つに絞り込めない。この場合、金メダリストというキーワードは曖昧であると考え、このキーワードに対して得られた限定表現のグループを提示する。これによりユーザが自分の質問の曖昧性に気付いていない場合でも、質問が曖昧であることを知ることができ、適切な解答を選択することができる。

- 曖昧な質問に対して複数の解答を表示する

2.2 節で述べたリスト型の質問応答システムは、「日本三景は何と何と何ですか」という質問のように1つの質問に対して複数の解答が得られる場合に限り有効であったのに対し、本研究では「ワールドカップの優勝国は何処ですか」という質問のように、質問が曖昧であるものに対応している点異なる。また、解答のリストを表示にする際、結果を分かりやすく伝えるために、解答を曖昧なキーワードの意味と共に表示することで、ユーザの誤認識を防ぐことができる。

- 曖昧な表現を列挙する

2.3 節で述べた対話型の質問応答システムのうち、本研究では特に徳江らが行った質問の曖昧性解消と関連が深い。徳江らは解答が複数得られた場合問い返しを行い、解答を1つに絞ることを試みた。しかしユーザが必ずしも1つの解答を求めているとは限らない。複数の解答をユーザが1度に欲しい場合、毎回聞き返しを行うのは手間がかかる。そこで、曖昧な質問が検出された場合には、聞き返しを行わず曖昧なキーワードの意味とそれに対する解答をリストとして提示する。本研究では曖昧性の検出率を向上させるために、パターンマッチによる限定表現の抽出は行わず、同一文中に存在する名詞のうちキーワードと関連のある単語を限定表現の候補とみなしすべて取り出す。そして属性の情報から限定表現のグループを作り、そのグループを出力することで、ユーザが適切な解答を選択できる。

第3章 提案システム

この章では、本研究の質問応答システムの全体の概要について述べる。

3.1 質問応答システムの概要

本研究では曖昧な質問に対応した質問応答システムの構築を行う。キーワードの意味が曖昧であるとき、その曖昧性を検出し、キーワードの個々の意味と共に解答を出力する。例えば「アテネオリンピックの柔道の金メダリストは誰ですか」という質問が挙げられる。ここで、柔道には「60キロ級」、「66キロ級」のように階級が複数存在するため、解答を1つに絞ることができない。そのため、図3.1のように各階級ごとに「野村忠広」「内柴正人」という解答のリストを出力し、ユーザに提示する。

質問：アテネ五輪の柔道の金メダリストは誰ですか？		
解答：	(1)	野村忠広 60キロ級の金メダリスト
	(2)	内柴正人 66キロ級の金メダリスト
	(3)	鈴木桂治 100キロ超級の金メダリスト
	(4)	谷亮子 48キロ級の金メダリスト
	(5)	谷本歩実 63キロ級の金メダリスト
	(6)	上野雅恵 70キロ級の金メダリスト
	(7)	阿武教子 78キロ級の金メダリスト
	(8)	塚田真希 78キロ超級の金メダリスト

図 3.1: 提案システムの理想的な出力例

3.2 処理の流れ

本システムの処理は以下の6つである．

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．

文書検索 質問文解析で得られたキーワードが含まれている文書を取得する．

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．

限定表現候補の抽出 解答候補を得られた文書を対象にキーワードと共起する名詞をすべて取り出し、限定表現の候補とみなす．

解答群の選択 限定表現の候補が持つ共通の属性を手がかりに、最終的にリストとして提示すべき解答群を生成する．

解答の出力 得られた限定表現と解答をリスト形式で提示する．

上記の処理を行う本研究の質問応答システムにおいて、従来の手法とほぼ同じ処理を行う質問文解析、文書検索、解答候補抽出の詳細を以下の節で述べ、本研究のテーマである限定表現候補の抽出、解答群の選定、解答の出力の詳細は4章で述べる．

3.2.1 質問文解析

ユーザからの質問文を解析し、以下の情報を取得する．

- キーワード K

入力された質問文の中から解答候補の手がかりとなるキーワードを抽出する．本研究では、キーワードには主に名詞を用いる．抽出したキーワードは以下の2つの処理に用いる．

- － 文書検索に用いるキーワード

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．

- － 解答候補抽出に用いるプライマリキーワード

解答候補はこのキーワードの近傍から取り出す．

質問文の中から解答候補抽出のための中心となるキーワードをプライマリキーワード k_1 とし、その他の名詞をセカンダリキーワード $k_i (i \neq 1)$ とする．それらのキーワード群を K とする．

- 解答タイプ *type*

入力質問文から解答を抽出するためには、質問文がどのような種類の解答を要求しているのか判断することが重要である。解答タイプとはこのような解答の種類を表す分類である。解答タイプはIREX[5]の固有表現タグに準じて、「人名」、「国名」、「地域名」、「組織名」、「時間」、「その他」の6種類に設定した。解答タイプの詳細を表3.1に示す。「その他」とはそれ以外の5種類の解答タイプにあてはまらないタイプであり、IREXの固有表現タグの中に対応するものが無いことを表す。

表 3.1: 解答タイプ

解答タイプの種類	解答タイプの概要
人名 [per]	「誰」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

- キーワードのタイプ

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3.2にキーワードタイプの一覧を示す。

- hyponym

プライマリキーワードと解答候補の間に「<解答候補>は<キーワード>です」のような上位-下位関係が存在するパターンである。例として「井上康生は金メダリストです」のような関係が挙げられる。

- agent

プライマリキーワードが動詞のとき、「<解答候補>が<キーワード>する」というように、解答候補がキーワードの動作主になる関係である。例として「井上康生が優勝する」のような関係があてはまる。

- other

解答候補とキーワードの間に上記のような関係がない場合である。

- 解析例

現状では、質問文の解析は自動では行っていない。これは、本研究の主題である曖昧性を検出するには、質問文解析の段階で不適切な解析結果が生じることは不都合であるからである。よって人手で質問文を解析した結果を現在のシステムの入力とし、

表 3.2: 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

質問文解析においては適切な解析が出来ているものとして取り扱う。例えば「柔道の金メダリストは誰ですか」という質問の場合、入力された質問の形態素情報から、「金メダリスト」、「柔道」というキーワードが抽出され、「誰ですか」という情報から解答タイプが「人名」と決定される。「金メダリストは誰」というパターンから「金メダリスト」と「誰」という間には上位-下位の関係が成り立っているため、キーワードのタイプが「hyponym」と決まり、また「金メダリスト」というキーワードがこの質問のプライマリキーワードとなることが分かる。

3.2.2 文書検索

質問文の解析によって抽出されたキーワードを入力として、条件に当てはまる文書を検索する。取得する文書はプライマリキーワード、セカンダリキーワードの AND 検索で全てのキーワードが出現する文書を全文検索により抽出する。なお、全文検索には suffix array[11] を用いた。また、入力にはオプションとして、文書検索単位と検索年を指定する。本研究では知識ベースとして、毎日新聞 1991 年から 2003 年までの新聞記事データベースを用いた。検索対象としては全国版の記事のみを扱い、文頭が同じ見出しの地方版は内容が重複しているため除外した。新聞記事は各記事毎に異なる ID がつけられている。文書検索単位とは文書を解析する際に、扱う記事の範囲を示す。文書検索単位は記事単位と段落単位の 2 種類があり、記事単位の場合記事全文を検索単位とし、段落単位では記事内の段落を検索単位とする。同じ記事内においても、段落が変わると話の内容が変わってしまうケースが多く存在するため、本論文で扱う文書は全て段落単位とした。また、検索年の指定によって何年の新聞記事を検索するか指定する。検索年は 1991 年から 2003 年の間の任意の期間を指定する。

ここで抽出された文書を解答候補検出に用いる。文書検索では質問入力の検索単位にマッチした範囲で全キーワードが出現した文書を抽出する。徳江らのシステムは動詞、形容詞などの活用形には対応していなかった。例えば「泳ぐ」という動詞をキーワードとした場合を例に挙げると、抽出された文書に「泳いだ」という動詞の活用形が存在しても、表層の文字列が異なるために「泳いだ」を抽出することはできない。そのような表記の揺れにも対応できるように、動詞、形容詞の転置インデックスを作成し、文書の抽出率の向上を図った。

3.2.3 解答候補抽出

文書検索で抽出された文書を基に解答候補を抽出する。検索された記事、キーワード、解答タイプ、キーワードタイプを入力とし、スコア付けされた解答候補とその文書が出力として得られる。解答候補の抽出の流れとしては以下の通りである。

- 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 Chasen[12] を、構文解析には Cabocha[10] を使う。

- 解答候補の抽出

抽出された文書から、以下の条件を満たす名詞を解答候補として抽出する。

- － 条件 A：形態素情報が解答タイプの条件を満たす。

ここで参照する形態素情報とは、

- * 固有表現タグ
- * 品詞タグ
- * カタカナ文字列であるか否か

の3種類である。品詞タグ付け及び固有表現タグ付けには「Cabocha」を用いた。例えば、解答タイプが「人名」のときは、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という条件のいずれかを満たす名詞を探す。表 3.3 には解答タイプ毎に満たすべき条件の一覧を示す。

- － 条件 B：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との関係を表すパターンをいくつか用意し、パターンにマッチする名詞は条件 B を満たすとして抽出する。これらの抽出パターンにマッチした単語を解答候補として取り出す。また、抽出のパターンを表 3.4 に示す。「→」は文節の係り先を表す。表 3.4 のスコアについては後述する。3.2.1 項で述べたキーワードのタイプによって用いるパターンが異なる。例

表 3.3: 解答候補が満たすべき条件

解答タイプ	NE タグ	品詞	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	

例えば「hyponym」タイプでは5つのパターンを適用する。「agent」タイプはプライマリキーワードが動詞のときであり、「< 解答候補> が→ < キーワード> する」というパターンを用いる。なお、表 3.4 の「近傍の名詞」はキーワードの近傍にある名詞を取り出すパターンを表す。これは、上記のパターンマッチに失敗したときに適用されるデフォルトのパターンである。したがって、キーワードのタイプが「hyponym」「agent」「other」の全てのときに適用される。

表 3.4: 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	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

- 解答候補へのスコア付け

解答候補は一般に複数得られるため、スコアを付けて優先順位を付ける。スコア付けは、

- － 構文パターン
プライマリキーワードと解答候補の関係を用いたスコア
- － 品詞パターン
解答候補の品詞情報を用いたスコア
- － 距離スコア
解答候補と各キーワードの距離を考慮したスコア

の3つの要素によって決定する。以下、解答スコアの算出式を式 (3.1) に示し、スコアの詳細について述べる。また、スコアの各要素の値は、人手で調整を行い決定した。

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

- － S_{pat} ・・・構文パターンによるスコア
- － w_{pat} ・・・構文パターンの重み

- s_{mor} ・・・品詞情報によるスコア
- w_{mor} ・・・品詞情報の重み
- s_{dis} ・・・距離情報によるスコア
- w_{dis} ・・・距離情報の重み

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 3.4 で示した。解答候補として適切なものが抽出できるパターンほどスコアが高く与えられている。また、 w_{pat} はスコアに対する重みを表しており、値は 0.4 としている。

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 Chasen, Cabocha 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。解答候補の形態素情報が「NE タグ」であった場合、もっともスコアが高くなり、次いで、「品詞情報」、「カタカナ文字列」の順にスコアがつけられる。例えば、解答タイプが「人名」を表す場合、解答候補が人名を表す < PERSON > タグがついていた場合「NE タグ」が適用され、「名詞-固有名詞-人名」という品詞情報の場合には「品詞情報」が適用される。それ以外では「カタカナ文字列」が適用される。解答候補の属性に対するスコアを表 3.5 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 としている。

表 3.5: 品詞情報のスコア

属性	スコア
NE 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。

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

- K …総キーワード数
- k_i …キーワード
- a_j …解答候補
- $dist(k_i, a_j)$ … k_i と a_j の距離
-

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

ここで、「 k_i と a_j の距離」とは、キーワードと解答候補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し、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になる可能性があるため、距離に常に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 $sent(k_i, a_j)$ 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としている。

なお、3.2.1節～3.2.3節で述べた処理は、下記の点を除いては徳江らのシステムと同じである。

- 転置インデックス

従来のシステムは動詞、形容詞などの活用形には対応していなかった。そのため、例えば「泳ぐ」という動詞をキーワードとしたとき、文書に「泳いだ」という動詞の活用形が存在しても、表層の文字列が異なるためにその文書を検索できなかった。そのような活用による表記の揺れにも対応できるように、動詞、形容詞の基本形の転置インデックスを作成し、文書の抽出率の向上を図った。

- 解答候補へのスコア付け

解答候補ごとのスコアの相対関係をより明確にするために、各スコアの値を0～1の範囲の値を取るように修正した。

第4章 解答リストの作成

本研究では質問文に含まれる曖昧なキーワードは1つであると仮定する．キーワードの曖昧性を検出する手法として，キーワードの意味を限定する単語を抽出し，それらを比較することで曖昧性の検出を行うことを提案する．ここでキーワードの意味を限定するような修飾語句や複合名詞を限定表現と呼ぶことにする．そして，同じキーワードが文書ごとに異なる限定表現を持てば，そのキーワードは意味的に曖昧であると判断する．

徳江らは限定表現を抽出するために以下のような4つのパターンを用意し，抽出パターンにマッチした限定表現を取り出すことを試みた．

- 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．例えば，「48 キロ級の金メダリスト」という表現があったとき，「金メダリスト」の限定表現 s_{no} は「48 キロ級」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．例えば，「女子柔道」という表現があったとき，「柔道」の限定表現 s_{prev} は「女子」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．例えば，「柔道 60 キロ級」という表現があったとき，「柔道」の限定表現 s_{succ} は「60 キロ級」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．例えば，「81 キロ級で金メダルを獲得した」という表現があったとき，「金メダル」の限定表現 s_{de} は「81 キロ級」となる．

しかし，抽出パターンが少ないために限定表現を獲得できず，曖昧性が検出できない場合が多々あった．例えば，「水泳の世界選手権で優勝したのは誰ですか？」という質問に対し，図 4.1 の文書から限定表現を取り出す場合を考える．質問のキーワードは「水泳」「世界選手権」「優勝」である．水泳には種目がたくさんあるので種目ごとの優勝者が存在する．従って，限定表現を抽出する必要があるが，パターンマッチを用いて抽出を行っても，こ

今年一月，豪州・パースで開かれた水泳の第六回世界選手権で，の大会から正式種目になった遠泳（二十五キロ）に5時間1分45秒78をマークして優勝した米国の大学生，チャド・ハンデビーさん（20）＝写真（左）＝が，育ての親で全米コーチの実績をもつ，リップ・ダールさん（56）＝同（右）＝とともに来日した．

図 4.1: パターンマッチで限定表現を抽出できない例

の例の限定表現である「遠泳」という単語は，パターンにマッチしないため抽出することができない．そこで本研究では，パターンに依存せずに限定表現を網羅的に抽出し，曖昧性の検出率を向上させる方法を提案する．なお 4.2 節までは「水泳の世界選手権で優勝したのは誰ですか？」という質問を例に挙げて手法を説明する．

4.1 限定表現候補の抽出

曖昧性を検出するためには，文書から確実に限定表現を抽出する必要がある．そこで本研究では，限定表現を取り出すために同一文中に存在する名詞のうちキーワードと関連のある単語を限定表現の候補とみなしすべて取り出す．名詞は文書解析で得られた記事を Chasen を用いて形態素解析することで抽出する．しかし，解析によって得られた名詞をそのまま利用する場合，不適切な候補も数多く含まれることが問題となる．そこで，キーワードと名詞間の関連度を計算し，関連度の高い名詞のみを限定表現の候補とみなす．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

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．このような条件を設定した理由は以下の通りである．

- Dice 係数の信頼性

例えば， $(X_1 = Y_1 = X_1 \cap Y_1 = 1)$ と $(X_2 = Y_2 = 1000, X_2 \cap Y_2 = 999)$ となる場合を考える．Dice 係数を $D_1(x_1, y_1)$ ， $D_2(x_2, y_2)$ とすると， $D_1(x_1, y_1) = 1$ ， $D_2(x_2, y_2) = 0.999$ が求まる．この 2 つを比較すると，出現する記事数が多い $D_2(x_2, y_2)$ より記事数の少ない $D_1(x_1, y_1)$ のほうが大きくなる現象が生じている．出現する記事数が少

ないほど *Dice* 係数の信頼度が低いため，頻度を ($X = Y = X \cap Y \geq 100$) に設定した．

- 処理の高速化

形態素解析で得られた名詞の組み合わせの数は莫大であり，すべての組み合わせをテーブルとして格納することができなかった．そこで，*Dice* 係数の低いものは不適切な名詞の組み合わせと判断し，それらを削除した．これは処理の高速化にも繋がっている．

限定表現を抽出する際に作成した *Dice* 係数のテーブルを用いる．キーワードと記事中の名詞について $D(x, y)$ を求め，ある一定の閾値以上のときに限定表現の候補として抽出した．本研究では閾値を 0.04 以上とした．また限定表現の候補を抽出する際，徳江らのパターンマッチも用いる．パターンマッチによって得られた限定表現の候補は信頼度が高いと考えることができる．更に限定表現の候補が複合名詞の場合，*Dice* 係数のテーブルに単語が記載されていない可能性がある．従って，単語が見つからなかった場合，単語の先頭一文字を削り，再び *Dice* 係数のテーブルを検索する．これを順次繰り返す．図 4.2 に「水泳の世界選手権で優勝したのは誰ですか」という質問に対する限定表現の抽出例を示す．キーワードは「*」で囲まれた名詞で「水泳」「世界選手権」「優勝」である．解答は四角で囲まれた「北島康介」である．限定表現の候補は下線が引かれている名詞であり，キーワードとの関連度が低かった「日本競泳陣」のような複合名詞は除かれている．

但し，抽出される限定表現はキーワードごとに異なる．詳細を表 4.1 に示す．表 4.1 は各キーワードごとに抽出された限定表現が示されている．

*水泳*の*世界選手権*は第 1 1 日の 2 3 日，当地で競泳 4 種目の決勝などを行い，男子二百メートルバタフライで山本貴司（近大職）が 1 分 5 5 秒 5 2 の日本新記録で銀メダルを獲得した．男子バタフライで日本選手が五輪・*世界選手権*でメダルを獲得したのは，1 9 5 6 年メルボルン五輪 2 位の石本隆以来，4 7 年ぶり．今大会の日本競泳陣のメダルは，男子百メートル平泳ぎ*優勝*の北島康介（東京 S C）に続いて 2 個目．

図 4.2: 限定表現候補の抽出例

ここまでの処理の結果，表 4.1 のような (a_j, k_i, s_k) という 3 つの組が大量に取得される．ここで s_k は限定表現の候補， k_i は限定表現 s_k に対するキーワード， a_j は s_k, k_i が抽出された文書における解答候補である．例えば，この文書から種目の曖昧性を表しているのは（北島康介，優勝，男子百メートル平泳ぎ）である．

表 4.1: (a_j, k_i, s_k) の 3 つ組の例

解答候補	キーワード	限定表現
北島康介	世界選手権	当地 競泳 4 種目 決勝 日本新記録 獲得 五輪 獲得 1 9 5 6 年メルボルン五輪 2 位 今大会
	水泳	競泳 4 種目 男子バタフライ 男子二百メートルバタフライ
	優勝	競泳 4 種目 決勝 日本新記録 獲得 日本選手 1 9 5 6 年メルボルン五輪 2 位 石本隆以来 4 7 年ぶり 1 9 5 6 年メルボルン五輪 2 位 今大会 男子百メートル平泳ぎ 2 個目

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できないと判断する。そこで、4.1 節で抽出された (a_j, k_i, s_k) の集合から、キーワード k_i が共通で、かつその限定表現 s_k が共通の属性 $attr$ を持つものを選別し、これを解答群 $AG(k_i, attr)$ 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4つがある。

- かぎ括弧で囲まれた表現

限定表現がかぎ括弧で括られている名詞を指す。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「エンデバー」』、『大河ドラマの「秀吉」』とあった場合、「エンデバー」、「秀吉」といった表現がこの属性を持つ。

- 末尾 N 文字

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、「男子バタフライ」は、末尾 1 文字が「E1:イ」、2 文字は「E2:ライ」、3 文字は「E3:フライ」という属性を持つことを表す。

- シソーラスによる意味クラス

限定表現の上位概念を属性とみなす。シソーラスには日本語語彙体系 [4] を使用した。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が多い。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、「獲得」という単語は「1868」の属性を持つ。日本語彙体系において「1868」の意味クラスは「取得」という意味クラスを表す。

- 数量表現+接尾語

数量表現と接尾語で構成されている限定表現が持つ属性。例えば、「男子百メートル平泳ぎ」という限定表現は、「< NUM > + メートル平泳ぎ」という属性を持つ。

表 4.1 で得られた限定表現 s_k の属性を表 4.2 に示す。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を表す。「φ」は属性が存在しないことを示す。

表 4.2: 各限定表現の属性

限定表現	N	E1	E2	E3	K	T
当地	$\varnothing$	地	当地	当地	$\varnothing$	0461
競泳 4 種目	数 + 4 種目	目	種目	4 種目	$\varnothing$	1680,2436
決勝	$\varnothing$	勝	決勝	決勝	$\varnothing$	1680,1756
日本新記録	$\varnothing$	録	記録	新記録	$\varnothing$	0919,1046,1113,1211,1490
獲得	$\varnothing$	得	獲得	獲得	$\varnothing$	1868
五輪	数 + 輪	輪	五輪	五輪	$\varnothing$	0710,1101,1233,1689
今大会	$\varnothing$	会	大会	今大会	$\varnothing$	1232,1233,1514,1515,1687,1689
男子バタフライ	$\varnothing$	イ	ライ	フライ	$\varnothing$	0549,1680
日本選手	$\varnothing$	手	選手	本選手	$\varnothing$	0251,0360
石本隆以来	$\varnothing$	来	以来	隆以来	$\varnothing$	2713
4 7 年ぶり	数 + 年ぶり	り	ぶり	年ぶり	$\varnothing$	0543,0842,2563,2695
2 個目	数 + 個目	目	個目	2 個目	$\varnothing$	$\varnothing$

4.3 解答群の作成

4.1 節で述べた (a_j, k_i, s_k) の集合と 4.2 節で述べた属性を基に、最終的にリストとして提示すべき解答群 $AG(k_i, attr)$ を作成する．ここでは考えられるすべてのキーワード k_i と属性 $attr$ の組み合わせについて $AG(k_i, attr)$ を生成し、それらを最終的にユーザに提示する解答群の候補とする．但し、解答が 1 つしかないグループは解答リストとしてふさわしくないので生成しない．質問が「アカデミー賞を受賞した人は誰ですか」のときに生成される解答群の候補の例として挙げる．アカデミー賞には主演男優賞や監督賞など部門に曖昧性がある．なお、質問文解析の結果キーワードは「優勝」「アカデミー賞」が抽出されたとする．このとき、表 4.3、表 4.4、表 4.5、表 4.6 のような解答群が生成される．

表 4.3 は、25 回、73 回のように「数 + 回」という属性を持つ限定表現で解答群を形成している．表 4.4 は主演男優賞、助演男優賞のように限定表現の「末尾 3 文字」が共通なもので解答群を形成している．表 4.5 は文字通り括弧で囲まれた解答群である．但し、解答数が 1 つなので実際にはこのような解答群は生成されない．表 4.6 は日本語語彙大系「1857」の意味クラス「褒賞」をもつ限定表現で解答群を形成している．

表 4.3: AG(アカデミー賞, 数 + 回) の解答群

解答候補	キーワード	限定表現
行定勲	アカデミー賞	2 5 回
グラディエーター アン・リー		7 3 回
頭山 マイケル・ムーア スピリテッド・アウェイ		7 5 回
ウィリアム・A・ウェルマン タイガー・ウッズ		1 回
シドニー・ポワチエ		1 0 回

表 4.4: AG(アカデミー賞, 男優賞) の解答群

解答候補	キーワード	限定表現
清兵衛	アカデミー賞	最優秀助演男優賞
シドニー・ポワチエ ヘストン ジャネット・マクティアー ニコラス・ケイジ ジュリア・ロバーツ アン・リー グラディエーター オン ハン トニー		主演男優賞
高倉健		最優秀主演男優賞
マイケル・ケイン キューバ・グッディング ジェームズ・コバーン マイケル・クラーク・ダンカン ウォーケン ジュリア		助演男優賞

表 4.5: AG(アカデミー賞, 括弧) の解答群

解答候補	キーワード	限定表現
J F K	受賞	ベストワン賞

表 4.6: AG(アカデミー賞, 1857) の解答群

解答候補	キーワード	限定表現
清兵衛	アカデミー賞	最優秀助演男優賞
ヘストン ジャネット・マクティアー ニコラス・ケイジ ジュリア・ロバーツ アン・リー グラディエーター オン ハン トニー		主演男優賞
ロバーツ ヒラリー・スワンク ジュリア・ロバーツ エミー ヒラリー アリス		主演女優賞
高倉健		最優秀主演男優賞
マイ カンヌ国際映画祭		大賞
ヒラリー・スワンク		アカデミー主演女優賞
中島健蔵 鳥井		93年日本芸術院賞
マーシャ・ゲイ・ハーデン アンジェリーナ・ジョリー		助演女優賞

4.4 解答群に対するスコア付け

一般に解答群 $AG(k_i, attr)$ は複数得られる. ここでは, これらの解答群の中から, 最終的にユーザに提示する解答群を 1 つ選択するために, 解答群にスコア付けを行う.

解答群 $AG(k_i, attr)$ に対するスコア $ScoreAG(k_i, attr)$ を式 (4.2) のように 4 つのサブスコア S_1, S_2, S_3, S_4 の重み付き和であると定義する.

$$Score(AG(k_i, attr)) = w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4 \quad (4.2)$$

$$S_1 = \frac{|AG|}{A}, S_2 = \frac{A_{type}}{|GA|}, S_4 = \frac{\sum_{a \in AG} s(a)}{|AG|} \quad (4.3)$$

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

解答候補数におけるスコア (S_1)

分母は解答候補の総数であり, 分子は解答群が持つ解答候補数である. これは, 多くの解答候補中に限定表現が出現することで解答毎の曖昧性を検出しやすくなるという考えである. 従って, スコアは解答群がたくさん解答を含むほど高くなる. 前節 4.3 の質問「アカデミー賞を受賞したのは誰ですか」を例に挙げる. 質問に対して得られた解答の総数は 100 個であったとすると, 表 4.3 の解答群 $AG(\text{アカデミー賞, 数+回})$ に存在する解答の数は 9 個であるのでスコアは 9/100 となる. また表 4.4 の解答群 $AG(\text{アカデミー賞, 男優賞})$ に存在する解答の数は 18 個であるのでスコアは 18/100 となり, 数量表現+接尾語の解答群より末尾 3 文字の解答群のほうがスコアが高くなる.

限定表現の異なり数によるスコア (S_2)

分母は解答群が持つ解答候補数, 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, 同じ解答候補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, 適切に限定表現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, 異なる限定表現が同じ解答候補に対して出現する場合に低いスコアを与える. 例えば, 表 4.7 のような解答群を考える. これは 4.2 節で挙げた質問で得られたグループの1つで, 末尾3文字の「女優賞」という属性を持つ. このグループには全部で解答が14個存在するが「ヒラリー・スワンク」が「主演女優賞」と「アカデミー主演女優賞」という2つの限定表現を持つので異なり数は13個となり, このグループのスコアは13/14となる. 一方, 同じ末尾3文字という属性を持つ表 4.4 の AG(アカデミー賞, 男優賞)を見ると, 解答候補は重複しないのでスコアは1となる.

表 4.7: 解答が複数の限定表現にまたがる例

解答候補	キーワード	限定表現
ロバーツ * ヒラリー・スワンク * ジュリア・ロバーツ エミー ヒラリー マイ ジャネット・マクティアー カンヌ国際映画祭 シドニー・ポワチエ オン アリス	アカデミー賞	主演女優賞
* ヒラリー・スワンク *		アカデミー主演女優賞
マーシャ・ゲイ・ハーデン アンジェリーナ・ジョリー		助演女優賞

属性のスコア (S_3)

$S(\text{attr})$ の式は基準となる限定表現の属性毎に与えられるスコアである. 表 4.8 に属性毎に与えられるスコアを示す. スコアは人手によって調整した. これはスコアが高い属性ほど限定表現間の共通性が高く, ユーザに提示する解答リストとして適切であるとみなしている.

表 4.8: 属性のスコア

属性	スコア
カッコ	1
数量表現 + 接尾語	1
意味クラス	0.4
末尾 3 文字	1
末尾 2 文字	0.5
末尾 1 文字	0.2

解答群中に存在する解答の平均スコア (S_4)

この式の分母は解答群が持つ解答候補数であり、分子は解答候補が持つスコアの平均である。この項は、限定表現が持つ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るほど解答群としてのスコアは高くなる。例を表 4.9 に示す。これは表 4.3 の解答群中の限定表現が持つ解答候補のスコアを示している。そして解答候補のスコアの総和を解答候補数で割ったものがグループのスコア、 $S_4=0.225$ となる。

表 4.9: 限定表現が持つ解答とスコアの例

解答	キーワード	限定表現	解答のスコア
行定勲	アカデミー賞	25 回	0.122
グラディエーター		73 回	0.114
アン・リー			0.118
頭山		75 回	0.134
マイケルムーア			0.770
スピリテッド・アウェイ			0.426
ウィリアム・A・ウェルマン		1 回	0.082
タイガー・ウッズ			0.129
シドニー・ポワチエ		10 回	0.129

全ての解答群について (4.2) 式のスコアを計算し、最も高いスコアを持つ解答群を 1 つ選択する。

4.5 解答の出力

一位に順位付けされた解答群を質問応答システムが出力する解答候補の信頼度のスコアが高い順にリストとして出力する．更に図 4.3 のように曖昧なキーワードの意味情報も出力する．現在は，各解答候補に共通のキーワードと限定表現を取得しているので，図 4.3 のようにそれを並べて出力する．図 4.3 のようにキーワードと限定表現を単に併記するだけでも，ユーザに自分の質問の曖昧性を認識させる効果があると考えられる．しかし，キーワードと限定表現から図 4.4 のような分かりやすい出力を生成することも重要である．ユーザにとって分かりやすい表現で解答リストを提示する手法を検討することは今後の課題である．

質問：アテネ五輪の柔道の金メダリストは誰ですか？
解答： (1) (野村忠広，金メダリスト，60キロ級)
(2) (内柴正人，金メダリスト，66キロ級)
(3) (鈴木桂治，金メダリスト，100キロ超級)
(4) (谷亮子，金メダリスト，48キロ級)
(5) (谷本歩実，金メダリスト，63キロ級)
(6) (上野雅恵，金メダリスト，70キロ級)
(7) (阿武教子，金メダリスト，78キロ級)
(8) (塚田真希，金メダリスト，78キロ超級)

図 4.3: 現時点のシステムの出力例

質問：アテネ五輪の柔道の金メダリストは誰ですか？
解答： (1) 野村忠広 60キロ級の金メダリスト
(2) 内柴正人 66キロ級の金メダリスト
(3) 鈴木桂治 100キロ超級の金メダリスト
(4) 谷亮子 48キロ級の金メダリスト
(5) 谷本歩実 63キロ級の金メダリスト
(6) 上野雅恵 70キロ級の金メダリスト
(7) 阿武教子 78キロ級の金メダリスト
(8) 塚田真希 78キロ超級の金メダリスト

図 4.4: システムの理想的な出力例

第5章 評価実験

5.1 実験方法

3,4章で述べた解答群を獲得する手法を評価するために、簡単な実験を行った。曖昧な質問を31個用意し解答群を抽出した。質問の一覧とそれに対する曖昧性を表5.2に示す。例えば、「ボクシングの世界チャンピオンは誰ですか」という質問は、「階級」と「日付」が曖昧であることを示している。現在、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表5.3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。キーワード、検索年、キーワードパターンの設定などは、何種類か試し、曖昧性がもっとも現れるような条件に設定した。4節の手法によって式(4.2)が最大となる解答群を選択した。そして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、正しい解答を含み、かつ得られた限定表現がキーワードの意味の曖昧性を正しく反映していることを指す。表5.1には質問の中でどのくらい適切な解答群が抽出できたのか示す。表5.1(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に現れた数である。(c)は適切な解答群は獲得できたが最上位に現れなかった数を表す。

表 5.1: 実験結果

適切な限定表現が存在しない (a)	8(26%)
適切な解答群が得られた	
適切な解答群が最上位 (b)	7(23%)
適切な解答群が最上位以外 (c)	16(51%)

5.2 考察

表5.1の(a)の結果より、適切な解答群が得られなかった質問の割合が約26%程度であることが分かった。これは限定表現の抽出方法に不備があることを示唆している。また、(b),(c)の結果より、限定表現が抽出された場合にそれが適切である割合は約74%程度であった。

表5.1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。

表 5.2: 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誰ですか	階級, 日付
2	サッカーの日本代表は誰ですか	性別, U-23, 日付
3	シドニー五輪の柔道の金メダリストは誰ですか	階級
4	百メートルの世界記録保持者は誰ですか	性別, スケート, 陸上, 水泳,
5	セリエAの日本人は誰ですか	サッカー, バレー, 日付
6	日本の首相は誰ですか	現在, 過去
7	日本シリーズで優勝したのは誰ですか	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誰ですか	NBA, NHL, 日付
9	棋聖戦で優勝したのは誰ですか	囲碁, 将棋, 日付
10	全英オープンで優勝したのは誰ですか	テニス, ゴルフ, 日付
11	全米オープンで優勝したのは誰ですか	テニス, ゴルフ, 日付
12	ワールドカップで優勝したのは誰ですか	サッカー, ラグビー, スキー, 日付
13	NHK杯で優勝したのは誰ですか	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誰ですか	F1, オートバイ, 日付
15	水泳の世界選手権で優勝したのは誰ですか	種目, 日付
16	グランドスラムを達成したのは誰ですか	テニス, ゴルフ, 囲碁, 日付
17	アカデミー賞を受賞したのは誰ですか	部門, 日付
18	レコード大賞を受賞したのは誰ですか	部門, 日付
19	ノーベル賞を受賞した人は誰ですか	部門, 日付
20	小泉内閣の官房長官は誰ですか	第何次
21	野球の新人王は誰ですか	チーム, 大リーグ, 日付
22	社民党の党首は誰ですか	日本, 外国
23	ジャイアンツの監督は誰ですか	読売, サンフランシスコ, 日付
24	タイガースの監督は誰ですか	阪神, デトロイト 日付
25	サリン事件がおきたのはいつですか	地下鉄サリン, 松本サリン
26	雪印が起こした事件はいつですか	食中毒, 牛肉偽装
27	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
28	高校野球で優勝したのは何処ですか	春, 夏, 日付
29	甲子園で優勝したのは何処ですか	春, 夏, 秋, 阪神, 日付
30	クラブユースサッカー選手権で優勝したのは?	U-18, U-15, 日付
31	世界貿易センターの高さは何メートルですか	大阪, ニューヨーク

表 5.3: 質問設定

質問 番号	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
1	世界チャンピオン	ボクシング	97-2003	per	hyponym
2	日本代表	サッカー	2000	per	hyponym
3	金メダリスト	シドニー五輪, 柔道	2000-2003	per	hyponym
4	世界記録保持者	百メートル	2000	per	hyponym
5	日本人	セリエ A	all	per	hyponym
6	首相	日本	2000	per	hyponym
7	優勝	日本シリーズ	2000	per	agent
8	優勝	西カンファレンス	2000-2002	per	agent
9	優勝	棋聖戦	2000-2002	per	agent
10	優勝	全英オープン	2000	per	hyponym
11	優勝	全米オープン	2000	per	hyponym
12	優勝	ワールドカップ	2000	per	hyponym
13	優勝	N H K 杯	2000-2002	per	agent
14	優勝	日本グランプリ	2002-2003	per	hyponym
15	優勝	世界選手権, 水泳	2002-2003	per	hyponym
16	達成	グランドスラム	2000	per	agent
17	受賞	アカデミー賞	2000-2003	per	hyponym
18	受賞	レコード大賞	2000-2003	per	hyponym
19	受賞	ノーベル	2000	per	hyponym
20	官房長官	小泉内閣	2000-2002	per	hyponym
21	新人王	野球	2002-2003	per	hyponym
22	党首	社民党	2000	per	hyponym
23	監督	ジャイアンツ	2000-2003	per	hyponym
24	監督	タイガース	2000-2003	per	hyponym
25	起きた	サリン事件	2000-2002	time	agent
26	事件	雪印	2000-2002	time	hyponym
27	優勝	六大学野球	2000-2002	org	hyponym
28	優勝	高校野球	2000-2003	org	hyponym
29	優勝	甲子園	2000	org	hyponym
30	優勝	クラブユースサッカー選手権	97-2003	org	hyponym
31	高さ	世界貿易センター	97-2003	ot	hyponym

解答群抽出成功例

図 5.1 に解答群が上手く抽出できたと判断した例を示す。これらの図で示したのは、キーワードと属性でまとめた解答群のスコアが上位のものである。図 5.1 では「ボクシングの世界チャンピオンは誰ですか」の問いを入力とし、得られた解答群のスコアが 1 位のグループを指す。また適切ではない解答群の例として 18 位のグループを示す。この図において、キーワードは「キーワード：世界チャンピオン」と表記されている部分に相当し、「世界チャンピオン」がキーワードである。また限定表現と解答候補は「スーパーフライ級：戸高秀樹」と表記されている部分に相当し、限定表現が「スーパーフライ級」で解答候補が「戸高秀樹」である。この場合、末尾 2 文字の「王者」という属性を持つ 18 位の解答群は、ボクシングの階級のグループを獲得できていないのに対し、末尾 3 文字の「ライ級」という属性を持つ 1 位の解答群は「スーパーフライ級」「フライ級」のようにボクシングの階級を示しているので正しい解答を持つ解答群として適切であると判断した。

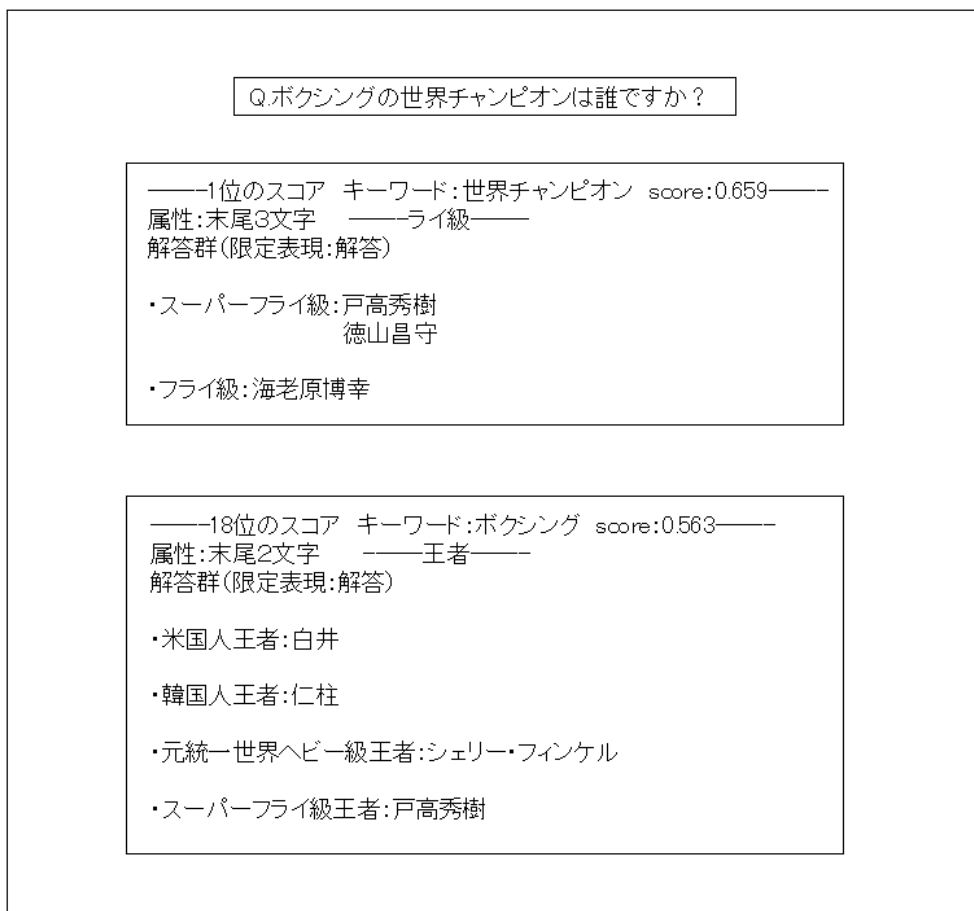


図 5.1: 正しい解答群の抽出例（ボクシング）

解答群は獲得できているが、順位付けが適切ではない例

図 5.2, 図 5.3, 図 5.4, 図 5.5, 図 5.6, 図 5.7, 図 5.8, 図 5.9 に適切な解答群が獲得できているにもかかわらず一位にならなかった例を示す。

図 5.2, 図 5.3 は「ノーベル賞を受賞したのは誰ですか」という質問を入力したときの結果である。それぞれの図は、ノーベル賞の部門に対する解答群を作成していて、正しい解答群である。正しい解答群が複数得られた場合、限定表現を多く含んでいる解答群のほうを選択すべきである。しかし、限定表現を多く含んでいる解答群のほうの順位が低くなっている。これはキーワードが「ノーベル」で末尾 3 文字の「理学賞」という属性を持つ解答群の方が、キーワードが「受賞」で末尾 2 文字の「学賞」という属性を持つものより、サブスコア S3 が高いので、末尾 3 文字の解答群が信頼度が高いと判断されているためである。正しい解答群が複数得られた場合の取り扱いは今後の課題である。

Q.ノーベル賞を受賞したのは誰ですか

——1位のスコア キーワード:ノーベル score:0.662——
属性:末尾3文字 ——理学賞——
解答群(限定表現:解答)

・物理学賞: 白川
 クロード・コーエンタンヌージ
 湯川秀樹
 江崎玲於奈
 白川英樹
 ストーマー

・生理学賞:利根川進

・医学生理学賞:ニール

図 5.2: 解答群が上位に現れない例（ノーベル賞の 1 位）

Q.ノーベル賞を受賞したのは誰ですか

——13位のスコア キーワード:受賞 score:0.601——

属性:末尾2文字 ——学賞——

解答群(限定表現:解答)

・物理学賞: 江崎玲於奈

ストーマー

白川

湯川秀樹

朝永振一郎

・経済学賞: セン

アマルティア・セン

・化学賞: 陳

白川英樹

A・セン英, 李遠哲

・文学賞: 高

高行健

大江健三郎

ダライ・ラマ

ラビドナース・タゴール

李遠哲

・生理学賞: 利根川進

・生理学医学賞: ジャック・モノ

図 5.3: 正しい解答群が上位に現れない例 (ノーベル賞の13位)

次に図 5.4, 図 5.5 は「シドニー五輪の柔道の金メダリストは誰ですか」の問いを入力とし, 得られたスコアが 1 位と 3 位の解答群である. 2 つのグループを比べると, 末尾 3 文字のグループの解答候補に重複するものがあるため, 異なり数に対するスコア (式 4.2 の S2) が数量表現 + 接尾語のグループより低くなっている. これにより正しいと判断すべき限定表現のグループが下位に表示されていることが分かる. ここで, 改めて図 5.5 の限定表現と解答候補のペアを見てみると, 図中の (a)(b)(c) に示すとおり, 1 つの解答候補が複数の限定表現を持つことが分かる. それぞれの解答には 2 つの限定表現が存在してはいるが, 限定表現の意味としては同じである. よって, 表現は異なるが同じ意味を持つ限定表現をうまく 1 つにまとめることができれば, 4.4 節で述べた S2 が高くなるため, 末尾 3 文字のグループが上位になることができると考えられる. そこで, 解答が複数の限定表現を持つ場合は限定表現同士を比較して, 全く同じ文字列が含まれていたなら文字列の長い方の限定表現を採用することを提案する. 例えば, 「48 キロ級」と「女子 48 キロ級」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」の方が長いので, 田村亮子の限定表現は「女子 48 キロ級」を選択する.

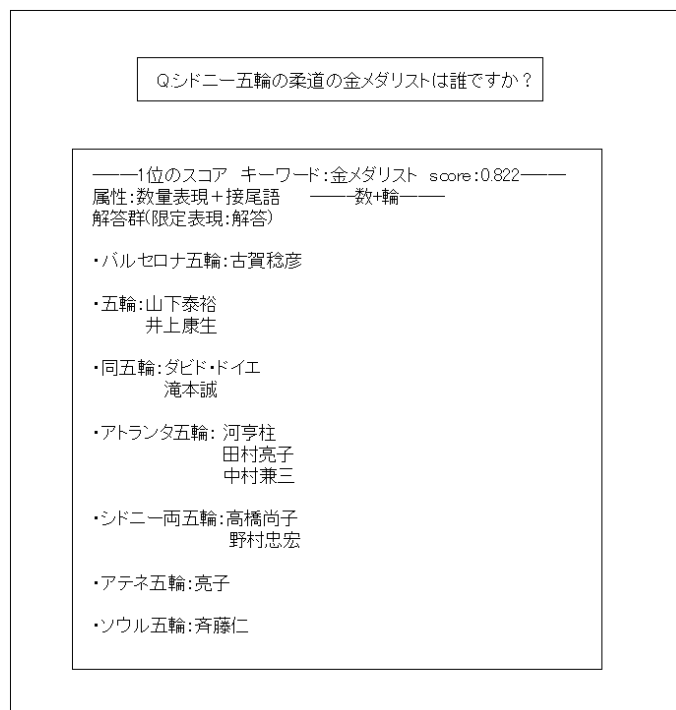


図 5.4: 正しい解答群が上位に現れない例 (柔道の 1 位のスコア)

Q.シドニー五輪の柔道の金メダリストは誰ですか？

——4位のスコア キーワード:金メダリスト score:0.771 ——
属性:末尾3文字——キロ級——
解答群(限定表現:解答)

・48キロ級: 上野雅恵
河亨柱
田村亮子

← (a)

・女子48キロ級: 田村亮子

・アトランタ五輪66キロ級: 敏仙

・女子70キロ級: 敏仙

・73キロ級: 滝本誠

・男子73キロ級: 中村兼三

・81キロ級: 武
武興
井上康生
滝本誠

← (b)

・男子81キロ級: 武興

・95キロ級: ケー・スンヒ

・100キロ級: 井上康生
野村忠宏
井上

← (c)

・男子100キロ級: 井上康生

図 5.5: 正しい解答群が上位に現れない例 (柔道の3位のスコア)

一方、図 5.6, 図 5.7 は「レコード大賞を受賞したのは誰ですか」という質問を入力したときの 1 位と 3 位の解答群である。2 つのグループを比較すると、サブスコア $S1$ のスコア図 5.6 よりも低かったため、順位に影響していることがわかった。この質問に関しては $S1$ の重みを高く設定すれば解答は上位に現れる可能性が高い。しかし汎用性を考慮すると、項の重み (w_1, w_2, w_3, w_4) をむやみに調節することは避けるべきである。そこで、式 (4.2) に以下の項を追加することを考える。

$$S5 = S_{type} \quad (5.1)$$

この項は限定表現の信頼度を表す。 S_{type} は 4 章で述べた限定表現を抽出するためのパターンに対するスコアである。パターンマッチによって得られた限定表現は、抽出率は低い信頼度が高いという特徴があった。そこで、4.1 節で述べた限定表現の抽出の段階で、限定表現の候補が抽出パターンによって抽出されたか否かという情報も付与し、スコア付けの段階で限定表現がパターン情報を持つなら $S_{type} = 1$ とし、持たなければ $S_{type} = 0.5$ として新たにスコア付けすることを提案する。ここで、図 5.6 と図 5.7 の限定表現が持つ $S5$ の情報を見てみると、図 5.6 中の限定表現は、パターン情報を持っていないので $S5 = 0.5$ となり、図 5.7 中の限定表現は、(b) がパターンの情報を持っていないながらも (a), (c) がパターン情報を持っているため $S5 = 1$ となる。これにより、正しい解答群に対するスコアが大きくなり、一位に順位付けされると期待できる。なお、従来のパターンマッチによる限定表現の抽出を行うと (a), (c) だけしか獲得できないのに対して、本研究の手法を用いると (b) も抽出できている。このことは、パターンに依存しない限定表現の抽出方法が有用であることを示唆している。

更に、図 5.8, 図 5.9 は「水泳の世界選手権で優勝したのは誰ですか」という質問を入力したときの 1 位と 19 位の解答群である。2 つの解答群を比較すると、サブスコア $S1$ が 1 位の限定表現のグループの方が高く、その他の項はほとんど同じであった。このような場合、それぞれの項の重み (w_1, w_2, w_3, w_4) を調節したとしても順位が入れ替わる可能性は少ない。そこで、先に述べた頻度と係りの情報である $S5$ の項を追加して再度スコアを計算すべきだが、この質問の場合、 S_{type} も同じであるために順位が変わることは無い。よって、このような質問に対する限定表現の抽出方法を新たに考えることは今後の課題である。

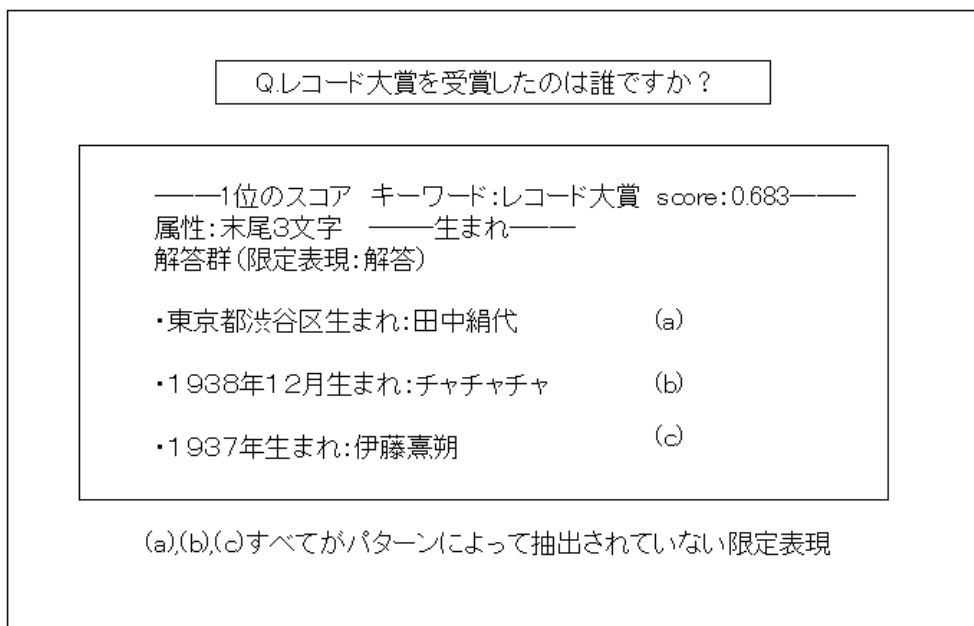


図 5.6: 正しい解答群が上位に現れない例 (レコード大賞の1位のスコア)

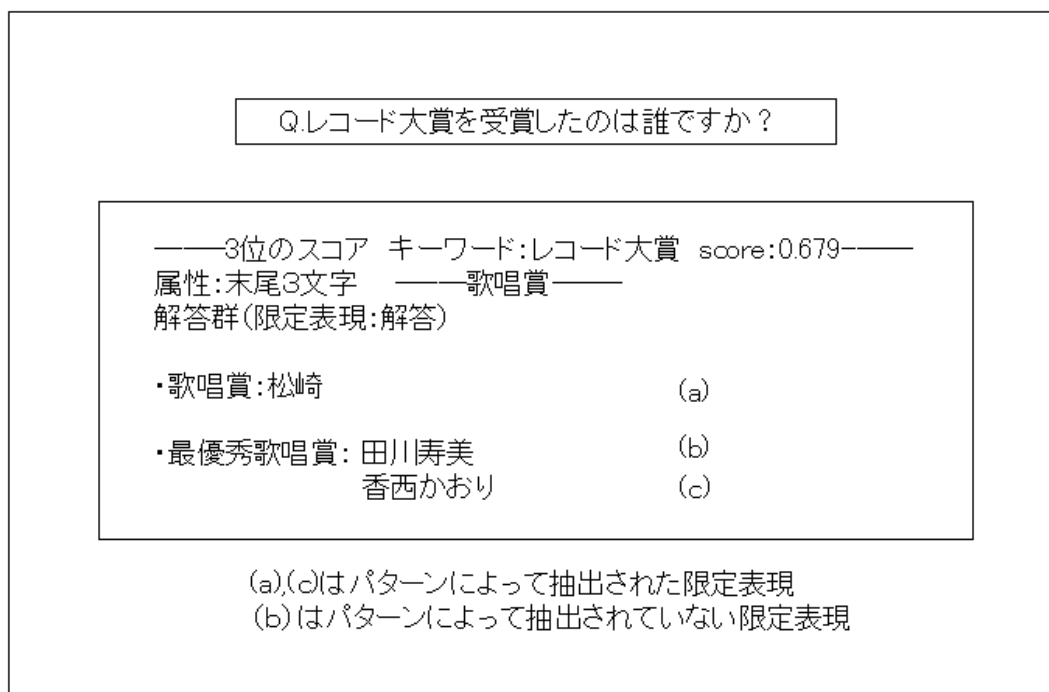


図 5.7: 正しい解答群が上位に現れない例 (レコード大賞の3位のスコア)

Q.水泳の世界選手権で優勝したのは誰ですか？

——1位のスコア キーワード:優勝 score:0.820——

属性:数量表現+接尾語 ——数+種目——

解答群(限定表現:解答群)

・決勝8種目:ピーター・ファンデンホーヘンバント

・飛び込み3種目:樋口まゆみ
小松

・8種目:萩原智子
セントラルイハシ
山野井智広

・12種目:藤野舞子
川口剛
山田沙知子

・競泳6種目:アレクサンドル・ポポフ
ロッカーク・イアン

・7種目:イアン・ソープ

・個人2種目:ミトリ・コモルニコフ

・3種目:武田美保
井村ク
ビルジニー・ドデュ
立花美哉

図 5.8: 正しい解答群が上位に現れない例 (水泳の1位のスコア)

Q 水泳の世界選手権で優勝したのは誰ですか？

——19位のスコア キーワード:優勝 score:0.712——
属性:末尾3文字 ——自由形——
解答群(限定表現:解答)

- ・同百メートル自由形: 川口剛
藤野舞子
- ・女子四百メートル自由形: イアン・ソープ
- ・女子八百メートル自由形: 石村元
- ・男子二百メートル自由形: ピーター・ファンデンホーヘンバント
- ・男子八百メートル自由形: グラント・ハケット
- ・男子五十メートル自由形: アレクサンドル・ポポフ
ロッキー・イアン
セントラルイハシ
山野井智広

図 5.9: 正しい解答群が上位に現れない例 (水泳の 17 位のスコア)

解答群抽出失敗例

31 個の質問中 8 個の質問は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．

- 曖昧な情報が記事に記載されていない

例えば「小泉内閣の官房長官は誰ですか」という質問がある．これは第 1 次，第 2 次，第 3 次小泉内閣という曖昧性があり，それぞれに官房長官が存在する．しかし，記事自体に「第 1 次小泉内閣の福田康夫官房長官」のような表記はされておらず，大抵は「福田康夫官房長官」のように解答候補とキーワードのみの情報が記載されている．従って第 1 次，第 2 次，第 3 次のような曖昧性が検出できなかった．

- 記事が存在しない

質問に対する解答が記載されている記事が少なく限定表現を抽出できなかった．記事の検索年数を「all」に設定すれば何とか限定表現を抽出できる可能性は高いが，Dice 係数のテーブルサイズが大きすぎてメモリ不足でプログラムを実行できなかった．しかし全く限定表現を取り出せなかったわけではなく，曖昧な限定表現は複数存在するため，そのうちのいくつかは抽出することに成功している．表 5.4 に抽出できた限定表現を示す．

表 5.4: 質問に対する正解の限定表現と抽出された限定表現のリスト

質問番号	理想的な限定表現のグループ	抽出された限定表現
8	NBA，NHL	NBA
9	囲碁，将棋	囲碁
14	F1，オートバイ	F1
24	デトロイト，阪神	デトロイト

- Dice 係数のテーブルに存在しない

4.1 節で述べたように限定表現として抽出する名詞は式 (4.1) の類似度が 0.0005 以上という条件を満たす必要がある．もし名詞と共起したキーワードが存在したとしても，その名詞の数が 99 個であったために限定表現として取り出されなかった可能性がある．更に名詞とキーワードの共起頻度が 0.04 よりも小さかったために，限定表現として取り出すことができなかったことも考えられる．よって共起頻度の閾値を新たに調節して，テストデータを用いて限定表現の抽出状況を調べる必要がある．

実験の結果より，曖昧な質問に対して本手法が有用であるとの見込みは得られた．そして得られた解答群を用いることで，曖昧な部分を問い返す文を作成し，対話的なやり取りができる質問応答システムに応用できる．しかし，正しい解答群が獲得できても解答が間

違っている場合が多々ある．今後，解答群の抽出精度を上げるには解答の抽出方法にも目を向ける必要がある．そしてより適切な実験設定で大規模な評価実験を行うことも大きな課題である．

第6章 結論

本論文では、ユーザから曖昧な質問が入力されたとき、得られた文書、解答候補から曖昧性を検出し、曖昧な情報と解答候補が一目で分かるように、それらをリスト形式で表示するリスト型質問応答システムについて述べた。本研究では特に曖昧性の検出率の向上に焦点を置き、以下に述べるような手法で曖昧性の解消を図った。

限定表現候補の出力

- キーワードと共起する名詞の抽出

限定表現を網羅的に抽出するために、まず文書を Chasen を用いて形態素解析し、名詞すべてを限定表現の候補とみなしすべて取り出した。取り出された名詞には限定表現の候補としてふさわしくないものも含まれている。よって抽出する名詞はキーワードと共起するものに限定した。そこで Dice 係数を用いて名詞間の共起性を定量化し、閾値以上の Dice 係数を持つ名詞のみを限定表現の候補として抽出した。

解答群の作成

- 属性によるグループ分け

数量表現 + 接尾語、末尾 1 文字、末尾 2 文字、末尾 3 文字、括弧、シソーラスという 6 つの属性を用いて解答群を作成した。

解答群の選定

- スコア付け

すべての解答候補と解答群内に存在する解答候補数の割合、限定表現と解答候補の異なり数、属性、解答群内にある解答の平均スコアという 4 つの観点からスコア付けを行った。正しい解答群が 1 位に現れたのは 31 個の質問中 7 個と少なかったが、正しい解答群は 23 個存在し、およそ 74 % の質問に対して正しい解答群を獲得することができた。

以上より、特定のパターンに依存せず限定表現を抽出する方法は有用であることが分かった。得られた解答群を用いて、問い返し文を生成することで対話的な質問応答システ

ムの構築に応用できると考えられる。しかし、正しい解答群が上位に来ていなかったり、グループ中に解答としてふさわしくないものが含まれていたりとは問題は山積である。それらを解決するために以下のような手法が考えられる。

- 解答候補抽出方法の改善

キーワードを含む文書を取り出してきてその中から解答を抽出する際、解答を間違っ
て取り出す場合が多々ある。これが解答群にふさわしくない解答が含まれる原因と
なっている。そこで、3章で述べた解答を抽出するためのパターンを増やし、正し
い解答の抽出率を向上させる必要がある。

- スコアの計算

5章でも触れたが、正しい解答群が上位に来るように、キーワードと限定表現の抽出
パターンの情報を用いたスコアをの導入を検討する。

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山田寛康 助手ならびに中村 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．

関連図書

- [1] Naoya Hidaka, Fumio Masui, Keiko Tosaki. Question Answering System Based on Expanded Answer Types and Multi-Scores. Proceedings of NTCIR-4, Tokyo, April 2003-June 2004.
- [2] Deepak Ravichandran, Eduard Hovy. Learning Surface Text Patterns for Question Answering System. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, July 2002, pp.41-47.
- [3] Jun'ichi Fukumoto, Tatsuhiko Niwa, Makoto Itogawa, Megumi Matsuda. Rits-QA: List answer detection and Context task with ellipses handling. Working Notes of the forth NTCIR Workshop Meeting pp.310-314 June 2004.
- [4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語彙体系. 岩波書店, 1997.
- [5] 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.
- [6] 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.
- [7] Joyce Y. Chai, Rong Jin. Discourse Structure for Context Question Answering. Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004, pp.23-30
- [8] 石下円香, 森辰則. 解候補スコアの分布を利用したリスト型質問応答. 言語処理学会第11回年次大会, D5-6, 2005.
- [9] 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-. 自然言語処理, Vol. 10, No.4, pp. 145-175, 2003.
- [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] Manber, U. and Myers, G. Suffix arrays: A new method for on-line string searches, 1st ACM-SIAM Symposium on Discrete Algorithms, pp. 319-327, 1990.

- [12] 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.