修 士 論 文

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院大学
情報科学研究科情報処理学専攻

徳江 英範

2005年3月

修 士 論 文

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

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

310073 徳江 英範

提出年月: 2005 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ついて述べる. このシステムは, ユーザの質問が曖昧であるときに, その曖昧性を解消するためにユーザに問い返しを行い, それに対するユーザの返答に基づいて最適な回答を選択する. ここでの曖昧な質問とは, 質問中の単語の意味が曖昧であるために解答を一つに絞ることができない質問を指す. 例えば,「ワールドカップの優勝国はどこですか」という質問は, ワールドカップにはサッカー, ラグビーなどの種類があるという意味で曖昧であり, これに対する解答を一意に決めることはできない. このような曖昧な質問が入力されたとき,「どの競技のワールドカップですか」といった問い合わせをシステムからユーザに行い, これに対するユーザからの返答によって適切な解答を決める. 本研究では, このような対話型質問応答システムの要素技術として, ユーザの質問の曖昧性を検出する手法について主に述べる. 提案手法の概要は以下の通りである。先の例で,「サッカーのワールドカップの優勝国はブラジル」と「ラグビーのワールドカップの優勝国はイギリス」のように,「ブラジル」「イギリス」と 2 つの解答候補が得られたとする. ここで, キーワードの意味を限定するような表現 (限定表現) に着目する. 例えば,「ワールドカップ」に連体修飾する「サッカーの」「ラグビーの」という句は「ワールドカップ」の意味を限定する表現とみなせる. このように, 質問文に含まれるキーワードについて, そのキーワードの限定表現を抽出し, 同じキーワードに対して解答候補毎に異なる限定表現が存在すれば, そのキーワードは曖昧であるとみなす.

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

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

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

次に, 質問の曖昧性の検出を行うために, 解答候補が抽出された文書からキーワードに係る表現を抽出する. 抽出する限定表現は, キーワードの「連体修飾句」,「格助詞デ格」で係る名詞, キーワードが複合名詞を作る場合その「直前の単語」,「直後の単語」, そして「日付表現」の 5 つである. 例えば, キーワードが「金メダリスト」のとき,「柔道の金メダリスト」という文からキーワードに連体修飾する「柔道」という単語が限定表現として抽出される. そして, キーワードと限定表現抽出の種類毎に限定表現のグループを作り, そのグループ内で個々の解答候補が異なる限定表現を持つかを調べることによって曖昧性の検出を行う. しかし, このグループ内には, 曖昧性を検出するには不要な限定表現も含まれることがよくある. そこで, 本研究では, 属性という概念を導入して不要な限定表現を取り除くことを試みた. 属性とは限定表現が持つ特徴のことを指す. 本研究では, 属性として「数

詞＋接尾語」「かぎ括弧」「意味クラス」「末尾 N 文字」「日付表現」の 5 種類を扱う. 例として,「５代目」「３代目」「水戸」という限定表現のグループがあった場合,「数詞＋接尾語」の属性では,「５代目」「３代目」が抽出され,「水戸」という限定表現は属性を持たない限定表現とみなし削除する. このように, 初期のと限定表現のグループ内で限定表現同士が共通の属性を持っていれば, それを改めて 1 つのグループとみなし, グループ内で属性を持たない限定表現を取り除く. 次に, このような一限定表現のグループは複数得られるため, これらに優先順位をつけるためのスコア付けを行い, 問い返しに最適なグループを 1 つ選択する. スコアは, 多くの解答候補で限定表現が現れているか否か, 属性を持つ限定表現の頻度, 異なる解答候補に対して同じ限定表現が得られているかどうか, などを考慮して決める. 本手法の評価を行うために, 曖昧性を含む質問 33 個に対して上記の手法で曖昧性の検出を行う実験を行った. その結果, 曖昧な質問であると判断された質問は全体の 80 %程度であり, そのうち, 適切な属性で限定表現が抽出された割合は約 80 %であった.

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の目的と背景

本論文はオープンドメインな対話型質問応答システムについて述べる. 一般的に, 質問応答システムでは, ユーザからの入力質問に対して的確な解答を返すことが課題となっている. 現在の質問応答システムの基本的な流れを以下に示す.

(1) 質問入力

(2) 質問文からのキーワード抽出

(3) 解答のタイプの選択

(4) キーワードによる文書検索

(5) 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, ユーザの入力質問に対して, 複数解答が得られる場合でも, その内の一つが正しければよいという方式になっている. そのため, このような一問一答型の質問応答システムでは, 個々のサブシステムを改良して, システム全体の精度をあげることを目的としている.

これに対して本研究では, 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. ユーザの質問が曖昧であるとは, 質問中の単語が曖昧であり, その単語の意味が明確でない状態を指す. 本論文は, このような対話型質問応答システムに必要な要素技術のうち, ユーザへの問い返し文の内容を決める方法を提案する. 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表現（限定表現）を抽出する. 次に, 同じキーワードに対して解答候補毎に異なる限定表現が存在すれば, そのキーワードは曖昧であるとみなす. このようにして検出された曖昧なキーワードの意味を尋ねる文がユーザへの問い返し文となる. 具体例として,「今年のノーベル賞受賞者は誰ですか.」という質問に対し, ノーベル賞 6 部門の受賞者がそれぞれ解答候補として選択されているならば, システムはこの質問は曖昧であると判定し,「何賞の受賞者ですか.」と問い返しを行うことで曖昧な質問に対応する. また, ドメインを限定しない曖昧な質問に対応可能な汎用的なシステムの構築を目指す.

1.2 本論文の構成

本論文の構成は以下の通りである.

- 2 章 関連研究

  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 対話型質問応答システム

  本論文のシステムの全体の構成について簡潔に述べる.

- 4 章 解答候補抽出

  従来の質問応答システムと同じ処理を行う質問文解析, 文書検索, 解答候補抽出について述べる.

- 5 章 曖昧性検出

  本論文の主題であるユーザの質問の曖昧性の検出, および問い返し文生成, 及び解答決定に関して述べる.

- 5 章 評価実験

  提案手法の評価実験とその考察について述べる.

- 6 章 結論

  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

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, 自然言語を用いた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. 質問応答システムとは, ユーザの入力した質問に対して, システムが解答となる単語や文章を提示するシステムである. 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　質問応答システムに関する研究

質問応答システムは 1980 年代に Wilensky[20] などのシステムが研究された. これらのシステムは, ユーザの意図が曖昧な場合に自然言語による問い返しを行う能力を備えていたが, そのためには人工言語で記述された, システムに特化した知識ベースが必要であった. しかし, 十分な能力を持つ人工言語の設計の困難さ, 知識ベース作成のコストなどの問題から, このような方法は明らかにスケーラビリティがない.

1990 年代になって, 電子化された大量の自然言語テキストが利用可能になったことから, 自然言語テキストを知識ベースとする質問応答システムの研究が盛んになってきた. 前田らは, 新聞記事 12 年分をデータベースとしたオープンドメインな質問応答システムを構築した [8]. システムの構成としては, 質問解析, 文書検索, 解答候補抽出, 解答選択, 根拠表示を行い, パッセージを表示するのではなく, 解答そのものを表示することが特徴となっている. 扱える質問は,「誰ですか」,「どこですか」のような解答が固有表現である場合と,「いつですか」,「何人ですか」,「何%ですか」といった数値表現にも対応している. 一方,「なぜですか」,「どのようにしてですか」などのような理由や方法に関する質問は取り扱っていない. また, このシステムは一問一答型のシステムであり, ユーザからの質問に曖昧な点は存在しないことが前提となっており, スコア上位 n 件を解答として出力し, 上位 m 件の中に解答が含まれているかという点で評価を行っている. Nomoto らは, キーワードタイプを判定するルールを 30 個作成し, また,29 種類の詳細な解答タイプで定義して,CLS(Common Longest Substring) によってスコア付けされた候補文書より, 解答タイプに一致するタグを解答として取り出す手法を試みた [13]. Takaki らは, 同じ解答を持つ文書が複数存在する場合に, トップスコア同士を比較する方法と全ての文書のスコアの合計で比較する方法の比較を行った [19]. Kawamura らは, 解答候補を含む文書から述語文法に着目して,「＜誰＞が＜何＞をした」などの関係を抽出する研究を行った [4]. Masui らは, 解答候補の抽出に, 係り受け解析, 構文木の深さ, 解答タイプの情報を使用してスコア付

けを行った [10]. Lee らは LSP（Lexico-Semantic Patterns）を用いて, 解答候補と解答タイプの関係をより詳細に分類した [16]. Ruck らは質問のタイプを人名だけに絞り, 抽出文書毎にスコア付けを行い, 上位文書から解答が存在するかどうかによって解答の順位をつけた [14]. Fukumoto らは, 質問からの解答タイプの決定にパタンマッチングを使用し, 解答候補を抽出するのにキーワードと解答候補の距離情報のみを用いた [3]. 現在, 質問応答システムに関しては,TREC や QAC などの評価型ワークショップなどで盛んに研究が行われている. 以前までは, これらのワークショップで扱われる質問応答システムは, ユーザの質問文には曖昧な点は存在しなく, 解答はただ一つに決まっているというスタイルであったが,「夏目漱石の名作は何ですか？」のように, 解答が一つではないような質問に対して評価を行うステージが追加され, 複数解答が存在するようなケースでの扱いに関しての研究も盛んに行われ始めている.

2.2 対話型質問応答システムに関する研究

対話型の質問応答システムに関する研究として, 黒橋らは京都大学の学術情報メディアセンターで使用されているヘルプシステムを構築した [7]. このシステムは自然言語で書かれた知識ベースとユーザ質問の柔軟なマッチングに基づいて, 曖昧な質問に対して自然言語による問い返しを行うことが出来るシステムである. データベースにはアプリケーション名などのような, キーワードとなるタイトルや, そのキーワードの定義, キーワードに関するマニュアルを記述したものを人手によって作成し, 質問の解答タイプの種類によって出力する項目を決める. 例として,「Mew でメールを送る方法は？」のような方法を聞く質問では,「Mew」というタイトルの中に存在するメールの送り方の記述を表示し,「Mew とは何ですか？」のような質問では,「Mew」タイトルの中に書かれているキーワード定義文を出力する. 質問の曖昧性を検出するために, 入力質問文とデータベースの症状に関する記述の類似度を計り, 複数の文書が候補として抽出された場合に問い返しを行う. 問い返し方法は, 複数の候補文書において異なる表現の個所を見つけ, その差異の単語をユーザに提示する. ただし, 知識ベースを人手で作成しているために, システムを適用可能なドメインは限定される.

清田らは, 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た対話型質問応答システムを提案した [5]. データベースにはマシンの症状とその対策が書かれている文書を用いている. ユーザへの問い返しを行うために, 頻繁に起こる質問に関しては, 対話カードを用いている. 対話カードとはユーザの質問に対してあらかじめ問い返し文を用意しておき, ユーザの質問が対話カードの文と一致した場合には, 決められた問い返しを行うものである. また, 対話カードに存在せずに, 複数の解答が得られた場合は, ユーザの質問に存在しない個所を抽出（状況説明文と呼ぶ）し, それらの一覧をユーザに提示することで曖昧性を解消している. 例を挙げると,「インターネットに接続出来ない」という質問に対し, 知識ベースに,「Windows98 でインターネットに接続できない」という文と「WindowsXP でインターネットに接続出来ない」という文書が抽出

された場合に, ユーザの質問である「インターネットに接続出来ない」という個所を除いた「Windows98」と「WindowsXP」の部分をユーザに提示して解答を選択させる手法である. このシステムで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ており, 扱う質問の範囲も Windows に関するドメインに限定される.

一方, オープンドメインな対話的質問応答システムの例として Sharon の研究がある [17][18]. 彼らが提案した HITIQA は「アルカイダの動向を調べる」といった情勢分析やレポート作成を支援するシステムであり, レポート作成に有用なパッセージをユーザに提示する. その際, ユーザとの対話によって提示するパッセージの内容を取捨選択する. 一方, ユーザとの対話を行わないが, ユーザの質問の曖昧性解消に関する研究として, 吉浦らの研究がある [21]. 彼らは, 多義語などによる曖昧性ではなく, 指示物決定の曖昧性などの, ユーザが想像しているイメージを把握するための手法を提案した. 著者らは, ユーザの質問を句単位にとらえ, それを深層格構造木に配置して, ユーザの意図するイメージを特定するためのクラスを工学的立場から 7 種類に分類した. このシステムは, ユーザへの問い返しを行うものではなく, ユーザによって与えられた文意を正しく解釈するための手法である.

2.3 本研究の特色

本研究の特色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ことが特徴である. 本論文で言う曖昧であるということは, 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意味している. 例として,「ノーベル賞の受賞者は誰ですか」といった質問が入力された場合, ノーベル賞にはノーベル医学賞やノーベル平和賞などがあり, 解答が一つに絞り込めない. この場合, ノーベル賞と言うキーワードに曖昧な部分が存在すると考え, このキーワードに対して「何のノーベル賞ですか？」という問い返しをすることが出来れば, 質問の曖昧性が解消できると考える.

- ドメインに依存しない

  2.2 節のように対話型の質問応答システムも存在する. しかし, これらのシステムはドメインに依存しているのに対し, 本論文では特定のドメインに特化せず, オープンドメインの知識ベースとして, 毎日新聞 1991 年から 2002 年までの記事を用いている点が特色である. 構造化されていないオープンドメインのテキストである新聞記事では, ドメインを限定したシステムのように, 頻繁に出現する曖昧な質問に対して, あらかじめ問い返し内容を登録しておくといった手法を行うには分野が広すぎるため困難である. 一方, ユーザの質問において, どの点が曖昧であるかについては, 解答を含むとして抽出された文書から判断することもできる. そこで, 本論文では質問文

中から得られたキーワードから条件にマッチした文書を抽出し, さらにその文書集合から解答候補を抽出できた文書内に存在するキーワードにかかる単語を比較することによって曖昧性を検出できると考える.

第3章　対話型質問応答システム

本研究で扱う対話の例を以下の表3.1に示す. ユーザが「ノーベル賞の受賞者」について質問を投げているが, いつの受賞者であるか, 何賞の受賞者であるかなどの曖昧性によって, システムは解答を一意に絞り込めない. そこでシステムがユーザに対し, 賞の種類や受賞の時期を聞き出すような質問をすれば, 適切な解答を絞り込むことができる. このようにユーザの質問の曖昧な個所を検出し, 問い返しを行うことが本研究の目的である.

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. システム全体の構成を図3.1に示す. ユーザからの質問文を入力として質問文解析, 文書検索, 解答候補抽出, 曖昧性検出, 問い返し, 解答決定の順にシステムは処理を行う. 以下に各プロセスの処理の概要を述べる.

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にマッチする文書を取得する. データベースとして毎日新聞1991年から2002年までを用いた.

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.

曖昧性検出 解答候補を得られた文書を対象にキーワードにかかる表現を抽出し, それらを比較することで曖昧性を検出する.

問い返し 曖昧性が検出された場合, ユーザに問い返しを行う.

解答決定 問い返しによってユーザから再度入力された情報にマッチする解答を提示する.

4章と5章にわたって, 本研究の質問応答システムの詳細について述べる.4章では, 本研究の質問応答システムにおいて, 従来とほぼ同じ処理を行う質問文解析, 文書検索, 解答抽出に関して述べる. 本研究の主題である曖昧性検出, 問い返し文生成, 解答決定に関しては5章で述べる.

表 3.1: 対話例

ユーザ「ノーベル賞を受賞したのは誰ですか」
↓
システム「どの賞の受賞者ですか」
↓
ユーザ「化学賞です」
↓
システム「何年の受賞者ですか」
↓
ユーザ「2002 年です」
↓
システム「2002 年にノーベル化学賞を受賞したのは田中耕一です」

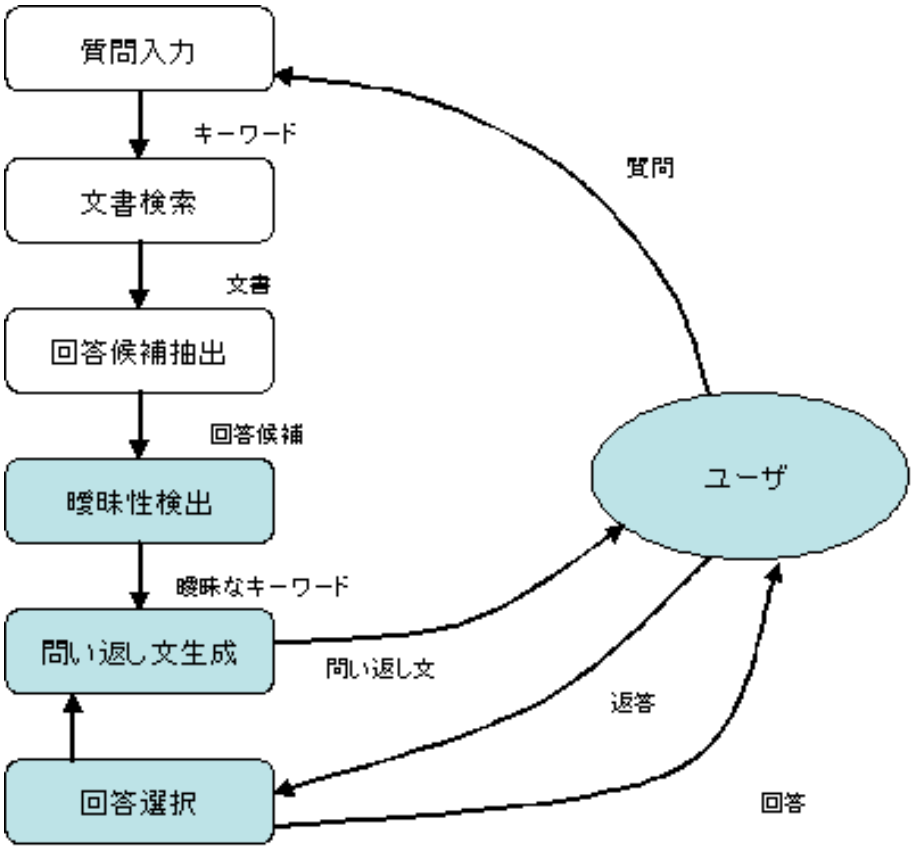

図 3.1: システム概要

第4章　解答候補抽出

4 章では, ユーザの質問から解答候補を抽出するまでの処理に関して述べる. 解答を抽出する手順は, 以下の通りである.

- 質問文の解析

  ユーザによって入力された質問文を解析し, キーワード, 解答タイプ, キーワードタイプの情報を得る.

- 文書検索

  質問文を解析して得られたキーワードを手がかりに解答候補を抽出するための文書を検索する.

- 解答候補抽出

  文書検索で抽出された文書群を対象に, キーワード, 解答タイプ, キーワードタイプなどの情報から解答候補を選択する.

以下の節では各モジュールについて述べる.

4.1　質問文解析

ユーザからの質問文を入力として, 以下の情報を出力として抽出する.

- キーワード K

  文書検索や解答抽出に用いるキーワード.

- 解答タイプ $type$

  質問に対する解答のタイプ.「誰」,「どこ」のようなキーワードによって決定する.

- キーワードのタイプ

  キーワードの中で中心となる主キーワードと解答候補との関係を示す.

4.1.1 キーワード

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.

- 文書検索に用いるキーワード

  キーワードによる全文検索のために用いるキーワード

- 解答候補抽出に用いる主キーワード

  質問文の中で解答候補を抽出する際に用いるキーワード. 解答候補は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（主キーワード）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.

4.1.2 解答タイプ

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. 回答タイプは IREX[2] 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 を種類設定した. 解答タイプの詳細を表 4.1 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は対応するものはない.

表 4.1: 解答タイプ

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

4.1.3 キーワードタイプ

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 4.2 にキーワードタイプの一覧を示す.

表 4.2: キーワードタイプ

名前	パタン	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	キーワードが動詞
other	その他	上記 2 パタン以外

- 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ンである. 例として「向井千秋は宇宙飛行士です」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「向井千秋が搭乗する」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.

4.1.4 解析例

現状では, 質問文の解析は自動では行っておらず, 人手で解析したものをシステムに入力している. 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ためである. したがって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. 図 4.1 では各項目の抽出の例を示す. 入力質問の形態素情報から,「受賞者」,「ノーベル賞」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.「受賞者は誰」というパタンから「受賞者」と「誰」という間には上位-下位の関係が成り立っているので, キーワードのタイプが上位-下位関係と決まり, また,「受賞者」というキーワードがこの質問の主キーワードとなることが分かる.

4.2 文書検索

質問解析によって出力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[9]

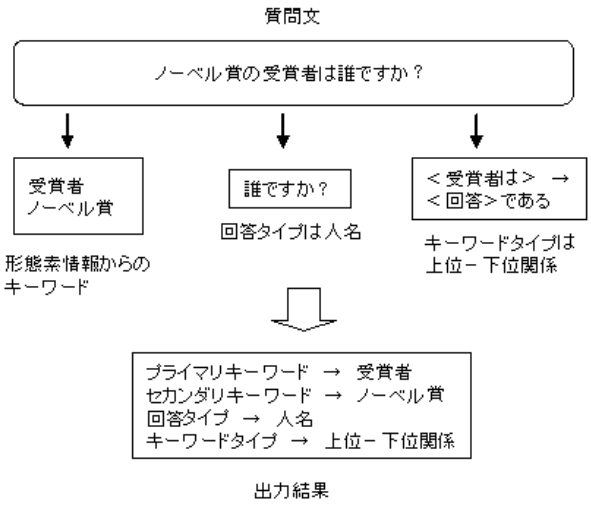

図 4.1: 質問解析例

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. 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. 図 4.2 の例で記事単位と段落単位の範囲を示す. 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 2002 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. 図 4.2 に文書検索の一連の処理を示す. 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. 本研究では知識ベースとして, 毎日新聞 1991 年から 2002 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する.

4.3 解答抽出

文書検索で抽出された文書を対象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.

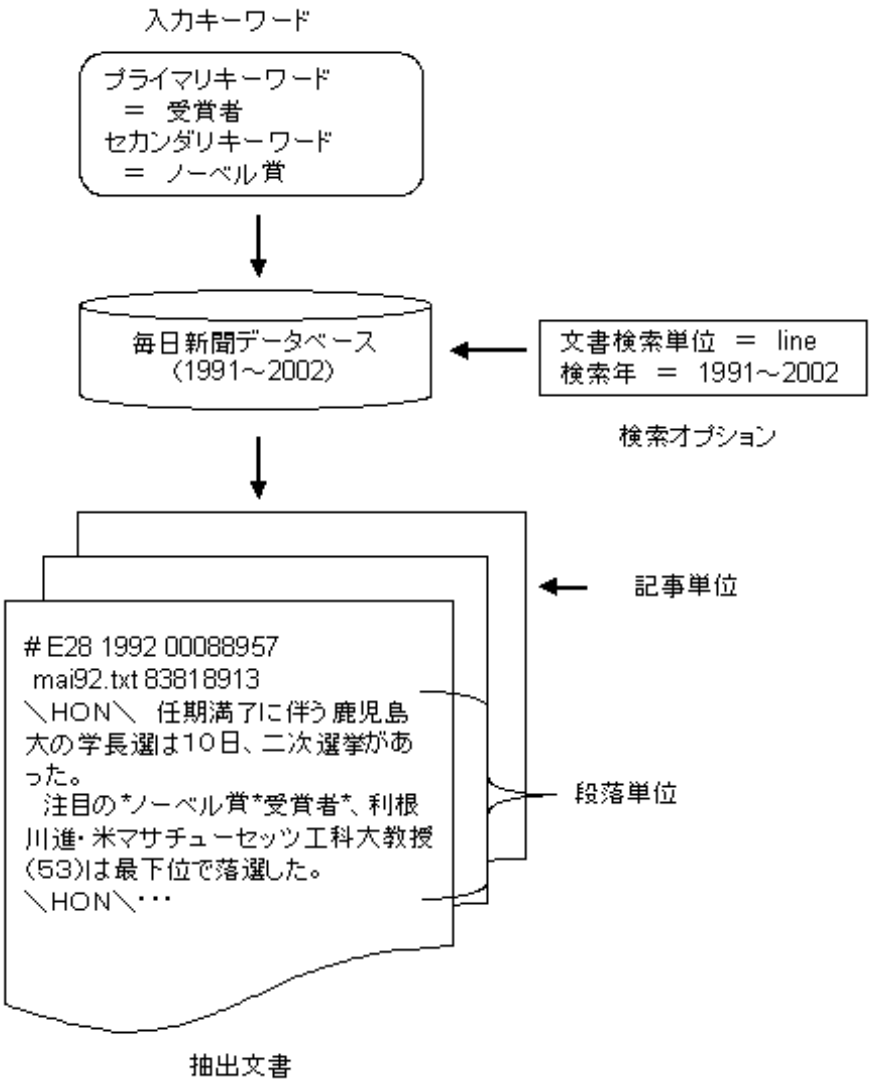
入力キーワード
プライマリキーワード
＝ 受賞者
セカンダリキーワード
＝ ノーベル賞
毎日新聞データベース
(1991～2002)
文書検索単位 ＝ line
検索年 ＝ 1991～2002
検索オプション
記事単位
E28 1992 00088957
mai92.txt 83818913
＼HON＼ 任期満了に伴う鹿児島
大の学長選は10日、二次選挙があ
った。
注目の*ノーベル賞*受賞者*、利根
川進・米マサチューセッツ工科大教授
(53)は最下位で落選した。
＼HON＼・・・
段落単位
抽出文書

図 4.2: 文書検索

- 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[11] を, 構文解析には Cabocha[6] を使う.

- 解答候補の抽出

  解析された記事の中から解答候補となる名詞を抽出する.

- 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.

解答候補の抽出ならびにスコア付けの詳細を以下の項で述べる.

4.3.1 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「南瓜」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 4.3 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：主キーワードの近傍にある.

  主キーワードと解答候補との関係を表すパタンをいくつか用意し, パタンにマッチする名詞は条件 B を満たすとして抽出する. 抽出のパタンを表 4.4 に示す.「→」は文節の係り先を表す. 表 4.4 のスコアについては後述する. 4.1.3 項で述べたキーワードのタイプによって用いるパタンが異なり, これらの抽出パタンにマッチした単語を解答候補として取り出す. 「hyponym」タイプでは 5 パタンと近傍の名詞を適用し,「agent」タイプはプライマリキーワードに動詞を割り当てるため, そのパタンも「＜解答候補＞が→＜キーワード＞する」というようなパタンおよび, 近傍の名詞を扱う.「other」タイプは近傍の名詞を抽出するだけのパタンである.

表 4.3: 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	

表 4.4: 解答候補抽出パタンとスコア

構文パタン	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	6
＜キーワード＞ハ→＜解答候補＞だ	hyponym	6
＜キーワード＞の→＜解答候補＞	hyponym	5
＜キーワード＞, ＜解答候補＞	hyponym	5
＜キーワード＞である＜解答候補＞	hyponym	5
＜解答候補＞ガ→＜キーワード＞する	agent	3
上記にないパタン（近傍の単語）	hyponym,agent,other	1

4.3.2 解答候補へのスコア付け

解答候補となる単語に対しスコア付けをし, 優先順位をつける. スコア付けは,

- 構文パタン

  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 (4.1) 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. 図 4.3 では解答候補へのスコア付けの例を示す.

$$S_{ans} = w_{pat} \times S_{pat} + w_{mor} \times S_{mor} + w_{dis} \times S_{dis} \tag{4.1}$$

- s_{pat}···構文パタンによるスコア
- w_{pat}···構文パタンの重み
- 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 4.4 で示した. 解答候補として適切なものが抽出できるパタンほどスコアが高く与えられている. 図 4.3 を例に挙げて説明する. 図の係り受け解析は解答候補を含む文章を「Cabocha」によって解析した結果を示す. これを見ると, 主キーワード「金メダリスト」と解答候補「田村亮子」の間には「＜キーワード＞の→＜解答候補＞」という関係が成り立っているので,「hyponym」タイプによる構文パタンのスコアは 5 となる. w_{pat} はスコアに対する重みであり, その値は 2 としている.

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

ここでは, 解答候補の品詞情報によるスコアについて述べる. 解答候補として抽出された単語の品詞情報を Cabocha によって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 4.5 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 1 としている. 図 4.3 を例に説明する. 固有表現抽出の図は Cabocha によって解析された出力であり, 左の列から,「解析された単語」,「単語の品詞」, 「固有表現タグ」を表している. 固有表現タグ付けの結果, 解答候補である「田村亮子」という単語には「＜ PERSON ＞タグ」がついているので, スコアは 3 となる. 品詞情報は NE タグ, 品詞, カタカナの順で優先される.

表 4.5: 品詞情報のスコア

属性	スコア
NE タグ	3
品詞	2
カタカナ	1

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 (4.2) 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.

$$S_{dist} = \frac{1}{K} \sum_{i=1}^{N} \{ \frac{1}{dist(k_j, a_i) + 1} \times sent(k_j, a_i) \} \tag{4.2}$$

- K···総キーワード数
- $dist(k_j, a_i)$··· 各キーワードと解答候補の距離
- $$sent(k_j, a_i) = \begin{cases} 1 & (\text{各キーワードに対して解答候補が一文中に存在しない場合}) \\ 3 & (\text{各キーワードに対して解答候補が一文中に存在する場合}) \end{cases}$$

ここで, キーワードと解答候補の距離とは, 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では, 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では, 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. 図 4.3 の例の距離測定の部分ではキーワードと解答候補の間の文字数が示されている. キーワード「金メダリスト」,「シドニー」, 「五輪」,「柔道」と解答候補「田村亮子」との間の距離はそれぞれ 1,17, 15,13 となる. 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. そのため, それぞれのキーワードについて, 一文中にキーワード, 解答候補が存在しないときにはさらにスコアを低くする. これは式 (4.2) の $sent(k_j, a_i)$ によって実現されている. 図 4.3 の例は, 解答候補と全てのキーワードが一文中に出現している例である.

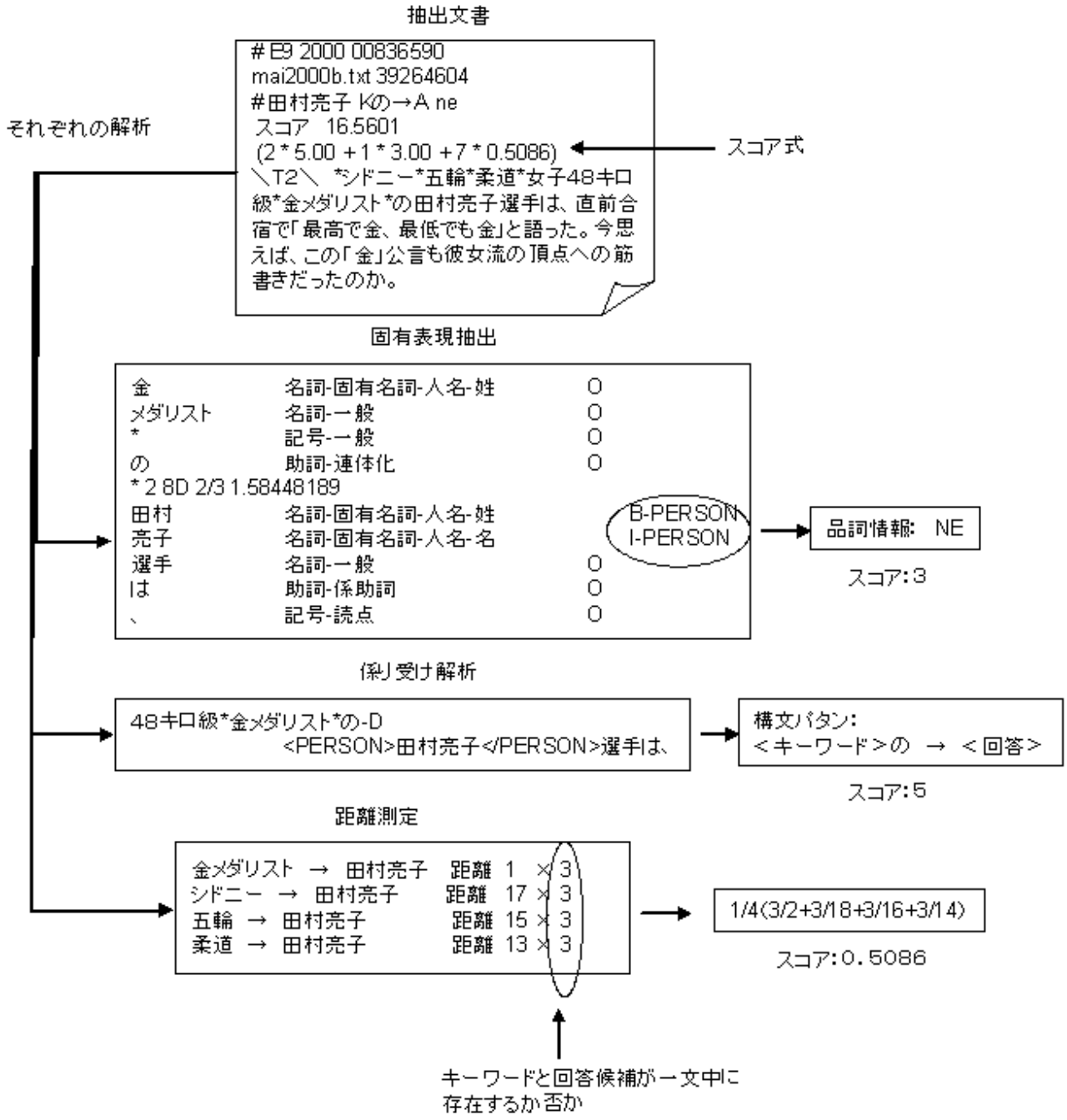

図 4.3: 解答候補スコア計算例

第5章　曖昧性検出

本研究では, 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えるのではないかと考えた. それぞれの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ここでは呼び, 限定表現を抽出するためのパタンをいくつか用意し, 抽出パタンにマッチした限定表現を取り出す. 取り出された限定表現はキーワードと抽出パタンのグループにまとめる. この際, 各グループには曖昧性を検出するために有用でない情報が混在しているので, それらを取り除くために, グループ毎の限定表現がどのような共通の属性を持つか検出する.

以下の節では曖昧性検出の詳細について述べていく.

5.1　限定表現抽出

本論文では, 曖昧なキーワードの意味を限定する修飾句や複合名詞に着目した. このような表現を限定表現と呼び, 各キーワードに存在する限定表現を比較することで曖昧性の検出を行った.

解答を得られた記事において, 記事中のキーワードを修飾する単語や, 複合名詞などを構成している単語を抽出する.Cabocha によって係り受け解析を行い, あらかじめ用意したパタンにマッチした場合に, マッチした単語を抽出する. 抽出するパタンは以下の 5 種類である.

- 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.

- 日付表現 (s_{date})

  キーワードの近辺に現われる日付表現．具体的には，固有表現タグとして “DATE” が付与され，キーワードの近辺にあり，かつ動詞の格要素となる表現を抽出する. 日付表現が月日を含む場合は年のみを取り出す．また,「今年」のような表現は記事の掲載日を参照して西暦年に正規化する．さらに，日付表現の主辞となる動詞も同時に抽出する．例えば,「2000 年 7 月に開催された世界大会」という表現があったとき，限定表現 S_{date} は [2000 年; 開催する] となる.

表 5.1 に限定表現の抽出パタンと例を示す.「→」は単語の係り先を表す. 「日付表現」は他のパタンと異なり, キーワードの意味を限定する要素ではなく, 解答に関するイベントがいつ起きたかを限定し, 他の 4 種類の限定表現とは意味合いが異なる. 日付表現の詳細は後述する.

表 5.1: 限定表現抽出パタン

	抽出パタン	例
連体修飾	＜限定表現＞の→＜キーワード＞	＜４８キロ級＞の＜金メダリスト＞
直前の単語	＜限定表現＞＜キーワード＞	＜５代目＞＜黄門＞
直後の単語	＜キーワード＞＜限定表現＞	＜大河ドラマ＞＜「信長」＞
デ格	＜限定表現＞で→＜キーワード＞	＜モーグル＞で＜金メダル＞

なお, 限定表現抽出パタンによって得られた限定表現のうち, 以下に該当する文字列は削除する.

- キーワードと同じ文字列
- 解答候補と同じ文字列

これは, 例として, キーワードに「金メダリスト」「柔道」ととした質問で, 「柔道の金メダリスト」という連体修飾パタンの限定表現が抽出された場合, キーワードとなっている単語自体が曖昧性解消の手がかりになることは無いと考えるためである. 解答候補の文字列を削除する理由も同様である. また, キーワードを含む複合名詞については, キーワード以外の個所を抽出して扱う. 例えば,「柔道男子６０キロ級」という限定表現があり, キーワードは「柔道」とあった場合, キーワードである「柔道」という単語を問い返しの際に入力されることはあり得ないため,「柔道」以外の単語の「男子６０キロ級」という表現

を取り出す. 同様に, 解答候補が限定表現に存在する場合も, 解答候補以外の文字列を限定表現として取り出す.

日付表現抽出

「サッカーのワールドカップの開催地はどこですか」という質問では, 年度によって開催された場所が違うという点に曖昧性がある. このような, 同じ名前のイベントでも年や日付によって生じる曖昧性に対して, 本研究では, 新聞記事の掲載日と記事内に存在する日付表現を比較してイベント発生の日時を特定し, 日付表現の曖昧性の解消を図った. 記事内の日付表現は, 固有表現抽出, および係り受け解析によって＜ DATE ＞の固有表現タグが付与された表現で, かつ以下の抽出パタンにマッチした場合に取り出す. 日付表現の抽出パタンを以下に示す. なお,「→」は文節の係り先を表す.

- ＜日付＞（に | で）→＜キーワード (動詞) ＞

  日付表現と助詞「に」「で」のいずれかが動詞となるキーワードに係るパタンを指す. 例としては, キーワードを「開催」とした場合,「2000 年に開催された」のような表現がこのパタンに適用される.

- ＜日付＞の→＜キーワード＞

  日付表現と助詞「の」がキーワードに係るパタンを指す. 例としては, キーワードを「ワールドカップ」とした場合,「2002 年のワールドカップ」のような表現がこのパタンに適用される.

- ＜キーワード＞（が | は | で）→＜動詞＞する
  ＜日付＞（に | から）→＜動詞＞する

  キーワードと助詞「が」「は」「で」のいずれかが動詞に係るパタンかつ, 日付表現と助詞「にから」が同じ動詞に係るパタンを指す. 例としては, キーワードに「サッカー」として,「2000 年にサッカーで優勝した」のような表現ではこのパタンが適用される.

- ＜日付＞（に | から） →＜動詞＞→（キーワード）

  日付表現と助詞「に」「から」がのいずれかが動詞に係り, さらに, その動詞がキーワードに係るパタンを指す. 例として, キーワードに「フランス」とした場合, 「2000 年に優勝したフランス」のような表現ではこのパタンが適用される.

抽出パタンの一覧を表 5.2 にまとめる.「日付表現抽出パタン」は一般的な抽出のパタンを表す.「2000 年にドラマの撮影を開始した」「2000 年にドラマが終了した」のように, 日付表現だけではその日時に何が起こったのか判断できないため, 日付表現を抽出する際には, そこにかかる動詞も抽出する. また, 表 5.2 の「動詞の有無」は抽出の際に動詞を抽出できるか否かを表す. また, 日付表現は「今年」や「先月」のように, 常に年月日が示され

ているわけではない. このような表現はあらかじめ日付表現を修正するリストを用意して対応した. 日時を表す表現のリストを表 5.3 に示す. また, 抽出記事内に正確な日時が表記されていることは稀であるため, 新聞記事の掲載日のリストを作成しておき, 表 5.3 と比較して日時を特定した. 図 5.1 に例を示す. 日時リストは左から掲載年, 記事 ID, 掲載月日を表しており, 図 5.1 の右の記事があった場合, リストから一致する ID を抽出し, 掲載日を特定する. 掲載日が特定出来たら, 記事中に存在する「去年」という表現より, 掲載年の 1991 年から 1 を引いた 1990 年と判断する. また, この際には, 表記されていない月日までは特定しない.

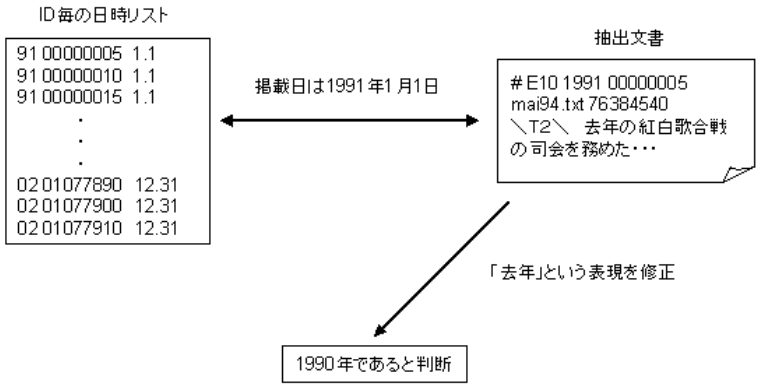

図 5.1: 日付表現の抽出例

表 5.2: 日付表現抽出パタン

日付表現抽出パタン	動詞の有無
＜日付＞＋助詞（に	で）→＜キーワード (動詞) ＞する	○
＜日付＞＋の →＜キーワード＞	×
＜キーワード＞＋（が	は	—で）→＜動詞＞する ＜日付＞＋（に	から）→＜動詞＞する	○
＜日付＞＋（に	から）→＜動詞＞した→＜キーワード＞	○

表 5.3: 日付表現リスト

日付を表す表現
今年, 去年, 昨年, おととし, 来年 再来年, 年始, 年末, 今月, 昨月 先月, 先々月, 来月, 今日, 昨日 おととい, 明日, 明後日, あさって

表 5.4: 限定表現の例

キーワード	解答候補	se_{no}(連体修飾)	se_{prev}(直前の名詞)	se_{succ}(直後の名詞)	se_{de}(デ格)
k_1 = 金メダリスト	a_1 = 田村亮子	４８キロ級 (1)	４８キロ級 (2), 女子４８キロ級 (1)		
	a_2 = 野村忠宏		６０キロ級 (1)		
	a_3 = 滝本誠			·(1)	
k_2 = シドニー五輪	a_1 = 田村亮子			４８キロ級 (1), 女子 (3), 代表選手·役員 (1)	
	a_2 = 野村忠宏				
	a_3 = 滝本誠			競泳 (1)	
k_3 = 柔道	a_1 = 田村亮子	女子マラソン (1)	女子 (4), コマツ女子 (1)	４８キロ級 (2), 女子４８キロ級 (1), 部 (1)	４８キロ級 (1)
	a_2 = 野村忠宏		男子 (1)	６０キロ級 (1)	
	a_3 = 滝本誠		男子 (1)	８０キロ級 (1)	

表 5.5: 日付表現

解答候補	日付表現
a_1 = 若田光一	2000.9 ＜＜９月＞＞；飛行 (3) 1996.1 ＜＜昨年 1 月＞＞；搭乗 (2) 2000.9.21 ＜＜9 月 21 日＞＞；打ち上げる (1)
a_2 = 毛利衛	1992.9 ＜＜昨年 9 月＞＞；搭乗 (2) 2000.2.12 ＜＜時間 12 日＞＞；打ち上げる (1) 2000.2 ＜＜今年 2 月＞＞；果たす (3)
a_3 = 向井千秋	1994.6.8 ＜＜八日＞＞；打ち上げる (1) 1998.10 ＜＜今年１０月＞＞；打ち上げる (2) 1994 ＜＜１９９４年＞＞；搭乗 (1)

5.2 限定表現のグループ化

抽出された限定表現は表 5.4 のようにキーワードとその限定表現のタイプによってグループ化される. 表 5.4 はキーワードと限定表現のタイプによって限定表現をまとめたものであり, 解答候補 a_1 から a_3 で抽出された限定表現を一つのグループとして扱う.

限定表現内の (1) などの数値は限定表現が抽出された数を表す. このように, 複数の解答候補を同じキーワードと限定表現のタイプでまとめ, それぞれのグループ内の限定表現を見ていくことで, どのキーワードに曖昧性が存在するかを検出する. また, 日付表現の例を表 5.5 に示す. これは解答候補に対して抽出された限定表現を示しており, 限定表現の左の数字が日付,「＜＜９月＞＞」などの表現は記事内に表記されている日付表現, 右が抽出された動詞を表し, 横の数字は限定表現の出現頻度である. 日付表現に関しては, 解答候補に関することがいつ起きたかという情報が曖昧な部分になっているため, キーワードによるグループ分けをせず, 解答候補 a_1 から a_3 で抽出された限定表現を一つのグループとして, 日付表現という限定表現の種類のみで曖昧性の検出を行う.

5.3 限定表現の属性

5.4 節において, 限定表現をキーワードと限定表現の種類によってグループ化した. しかし, キーワードと限定表現の種類だけによるグループ化のみでは, 曖昧性を検出する際に, キーワードの意味を限定するものではない表現が混在する可能性がある. 例として, 表 5.4 のキーワード「柔道」限定表現のタイプ「se_{succ}」の限定表現の中で「部」という単語は, キーワードの意味を限定するものではない. このような限定表現は問い返しの際にノイズとなる表現であるため取り除いていきたい. そこで, 本論文では, 各グループ毎の限定表現を属性 ($attr$) でまとめることで, 不必要な限定表現を取り除くことにする. 属性とは, 限定表現が持っている単語の特徴のことを示す. 属性導入の目的としては以下の通りである.

- 各グループ内に存在する余計な限定表現を省く

  キーワードと限定表現のタイプのグループを作る場合, 多くの場合, 曖昧性を解消するために有用でない情報が混じってしまう. このような情報は曖昧性の検出を正確に行うためには余計な情報であり, 出来るだけ省くことが重要である. そこで, 限定表現を属性でまとめることで, 属性を持たない限定表現を省き, ノイズを取り除くことを試みる.

- 限定表現の共通の属性を見つけることで問い返しを行いやすくする

  キーワードが曖昧性を持つ場合に, どのような属性を持っているかを明確にする. 共通の属性を持っているということは, 問い返しの際に, 属性について聞き返すことで多くの限定表現が抽出された場合でも, 自然な問い返しを行うことができる.

ここで用いる属性は以下の 5 つである.

- かぎ括弧で囲まれた表現
- 末尾 n 文字
- シソーラスによる意味クラス
- 数詞+接尾語の関係にあるパタン
- 日付表現によるパタン

この中で,「時間表現によるパタン」のみは, キーワードの曖昧性を判定するものではなく, 解答候補がいつ何をしたのかを検出するための属性なので, 他の属性と扱いが異なる.
以下,5 つの属性の詳細について述べる.

かぎ括弧

限定表現がかぎ括弧で括られている表現を抽出する. 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. 例として,「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 文字をそれぞれ抽出する. 例として,「４８キロ級」とあった場合, 末尾 1 文字は「E1:級」,2 文字は「E2:ロ級」,3 文字は「E3:キロ級」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抽出する. シソーラスには日本語彙体系 [1] を使用した. また, 上位概念としてある程度抽象度の高いものを用いる. さらに, 人名などの固有名詞のカテゴリに関しては問い返しを行う参考にはならないので, このようなカテゴリは省く. また, 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. 例として,「女子」という単語は「0046」の属性を持つ. 日本語彙体系において「0046」の意味クラスは「人間＜生物学的特徴＞」という意味クラスを表す.

数詞＋接尾語

数詞と接尾語で構成されているパタンの抽出を行う. 例えば, 「４８キロ級」という限定表現があった場合には,「４８＜NUM＞＋キロ級」という属性として扱う. また,「男子６０キロ級」などのような限定表現では, 数詞より前に存在する単語である「男子」の部分は削除し, 「６０＜NUM＞＋キロ級」を属性として扱う.「ドラゴンクエスト９」などのように数詞で終わるパタンでは,「＜NUM＞」を属性として扱い, 数詞の後に続く接尾語がないという情報を表す. この場合は「ドラゴンクエスト８」と「ドラゴンクエスト９」があった場合, 接尾語がないという点が共通として見出せる. 表 5.6 に数詞とみなす単語の一覧を示す.

表 5.6: 数詞として扱う単語の一覧

数詞として扱う単語
１, ２, ３, ４, ５, ６, ７, ８, ９, ０, 一, 二, 三, 四, 五, 六, 七, 八, 九, ○, 十, 百, 千, 万, 億, 兆, 初

日付表現

日付表現では, 日付表現と共に抽出した動詞を属性として扱う. 日付表現として抽出された限定表現は他の属性を持つことはない. 例を示すと,1999 年 5 月の記事に,「＜今月＞に＜開催される＞」とあった場合, 「今月」という表現を「1999 年 5 月」に修正し,「開催される」という動詞の基本形「開催する」を属性として持たせる.

上記の条件によって $S(k_x, type, attr)$ のグループに分けることができる. $S(k_x, type, attr)$ とは, キーワード k_x, 限定表現抽出パタン $type(no, prev, succ, de, date)$, それに属性 $attr$ を

持つ限定表現の集合を示す. 表 5.4 より例を示すと,
$S(k_3, succ, <NUM> +$キロ級$) =$ { ４８キロ級, ６０キロ級, ８１キロ級 }
$S(k_1, prev, E3 :$キロ級$) =$ { ４８キロ級, 女子４８キロ級, ６０キロ級 }
$S(k_3, prev, 0046) =$ { 女子, 男子, 男子 }
のようになる. また, 日付表現については, 表 5.5 を例に挙げると, $S(date$ ： 打ち上げる$) = \{2000, 1998, 1994\}$ のようにキーワード毎による区別がない.

図 5.2, 図 5.3, 図 5.4 には属性毎に限定表現のグループをまとめた例を示す. 左半分の図はキーワードと限定表現のタイプで抽出された解答候補 5 件とその限定表現を示し, 右の図は属性によってさらにまとめられた解答候補と限定表現のグループである.a_i は解答候補を表し, s は限定表現を表す. 限定表現の横の数字は出現頻度である. 意味クラスでは地名に関する情報に注目すると「ベンガルトラ」や「職場」という情報は,「どの地名か」という曖昧性を検出するには不必要な情報である. 上記の手法を適用することで地名のみが抽出され, 無駄な情報が取り除くことが出来ていることがわかる. その他のパタンにおいても共通の属性が取り出され, ノイズに当たると思われる情報を取り除いていることが分かる.

5.4 限定表現のグループに対するスコア付け

一般にグループ化された限定表現は複数得られる. 属性という概念は, 問い返し内容を決める要素でもある. どのような属性で問い返すのかを決定するために, キーワード k_x, 限定表現タイプ $type$, 属性 $attr$ よって決まるグループ $S(k_x, type, attr)$ にスコアを付け, 優先順位を決める. 限定表現のスコアを式 (5.1) に示す.

$$\begin{aligned} Score(S(k_x, type, attr)) \quad = \quad & w_1 \times \frac{A_{attr}}{A} + w_2 \times \frac{NT(S(\cdot))}{S(\cdot)} + w_3 \times \frac{\sum_{s \in S(\cdot)} O(s)}{\sum_{attr} \sum_{s \in S(\cdot)} O(s)} \\ & \times P(attr) \times \frac{\sum_{s \in S(\cdot)} O(s)}{S(\cdot)} \end{aligned} \tag{5.1}$$

- $A \cdots$扱う解答候補の数
- $A_{attr} \cdots$限定表現が属性 attr を持つ解答候補の数
- $NT(S(\cdot)) \cdots$属性を持つ限定表現の異なり数
- $SE(\cdot) \cdots$属性を持つ限定表現の数
- $O(s) \cdots$限定表現の出現頻度
- $w_1, w_2, w_3 \cdots$それぞれのスコアの重み

$(\cdot)$ は $(k_x, type, attr)$ の略を表す.

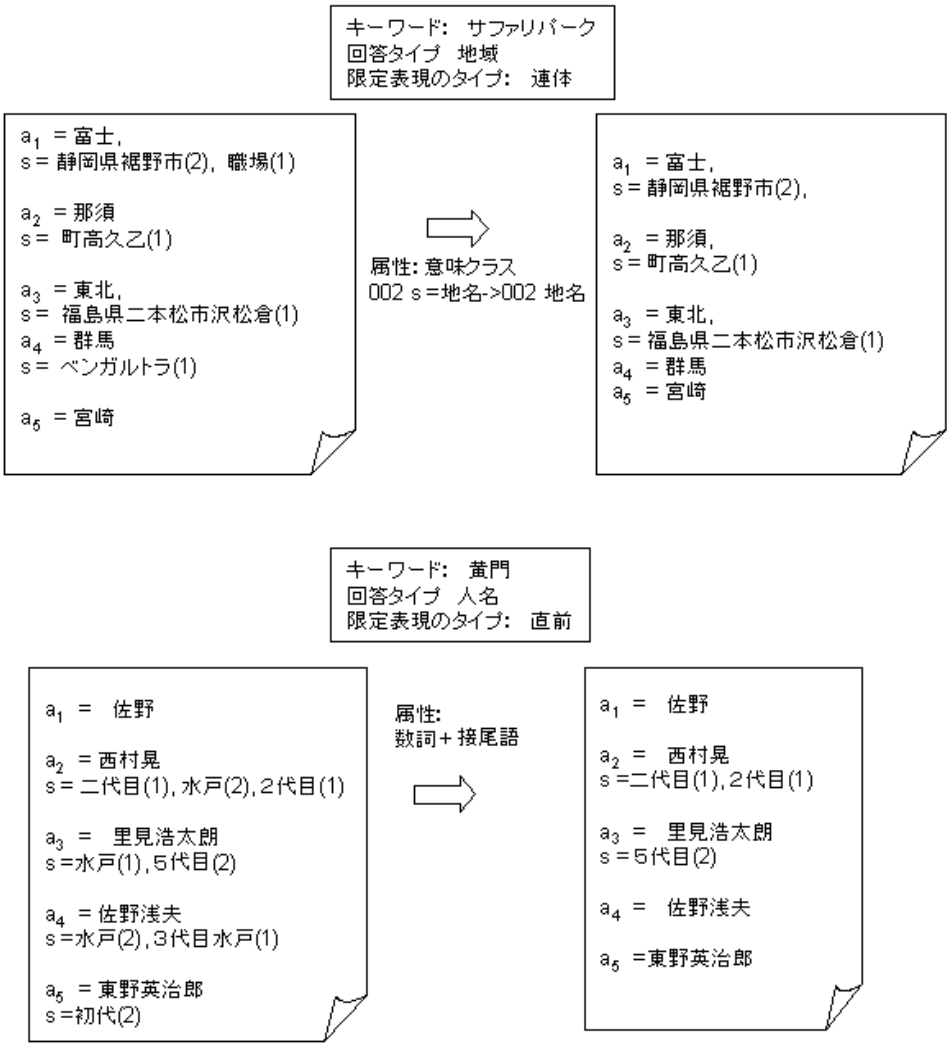

図 5.2: 属性でまとめた限定表現の例 1

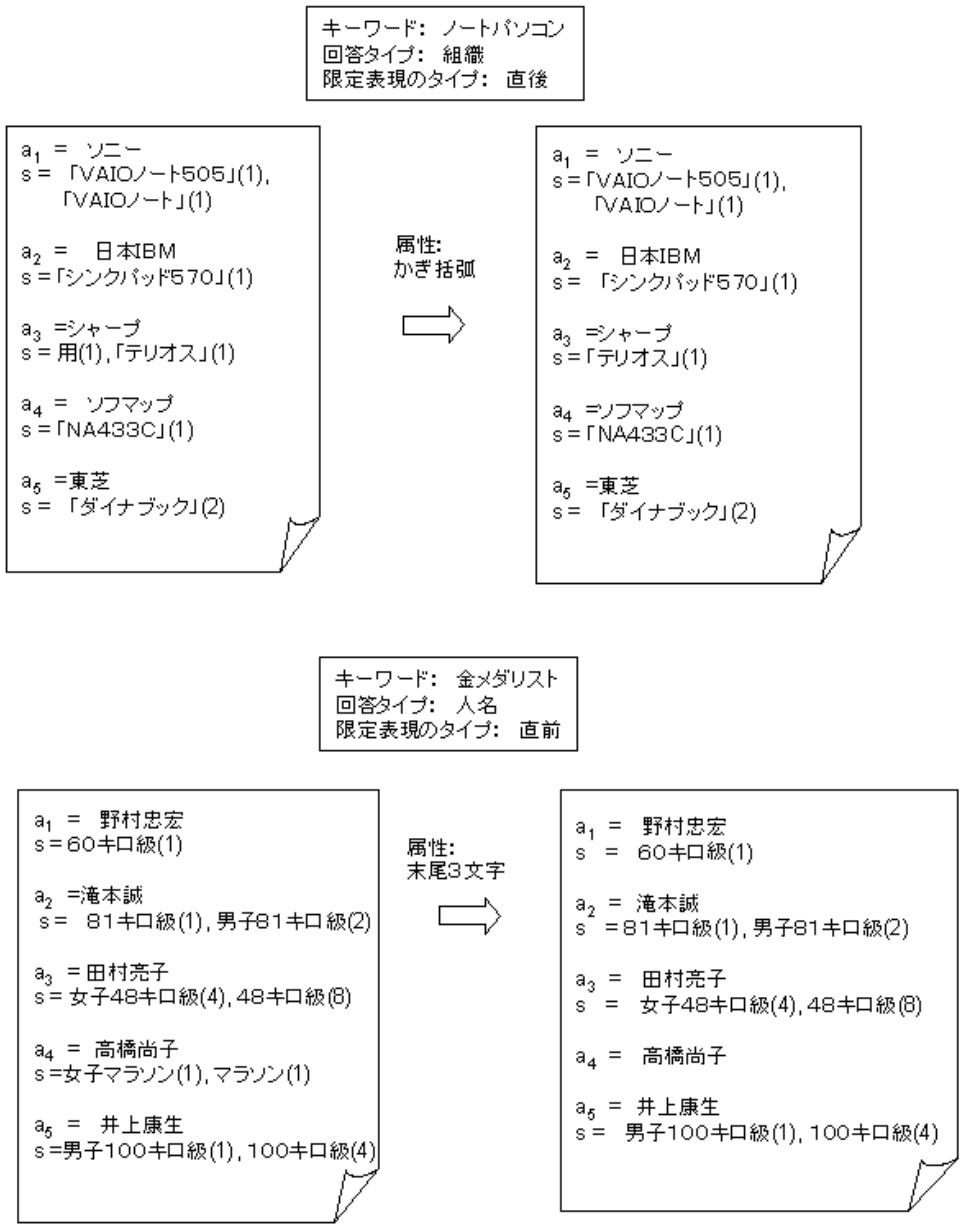

図 5.3: 属性でまとめた限定表現の例 2

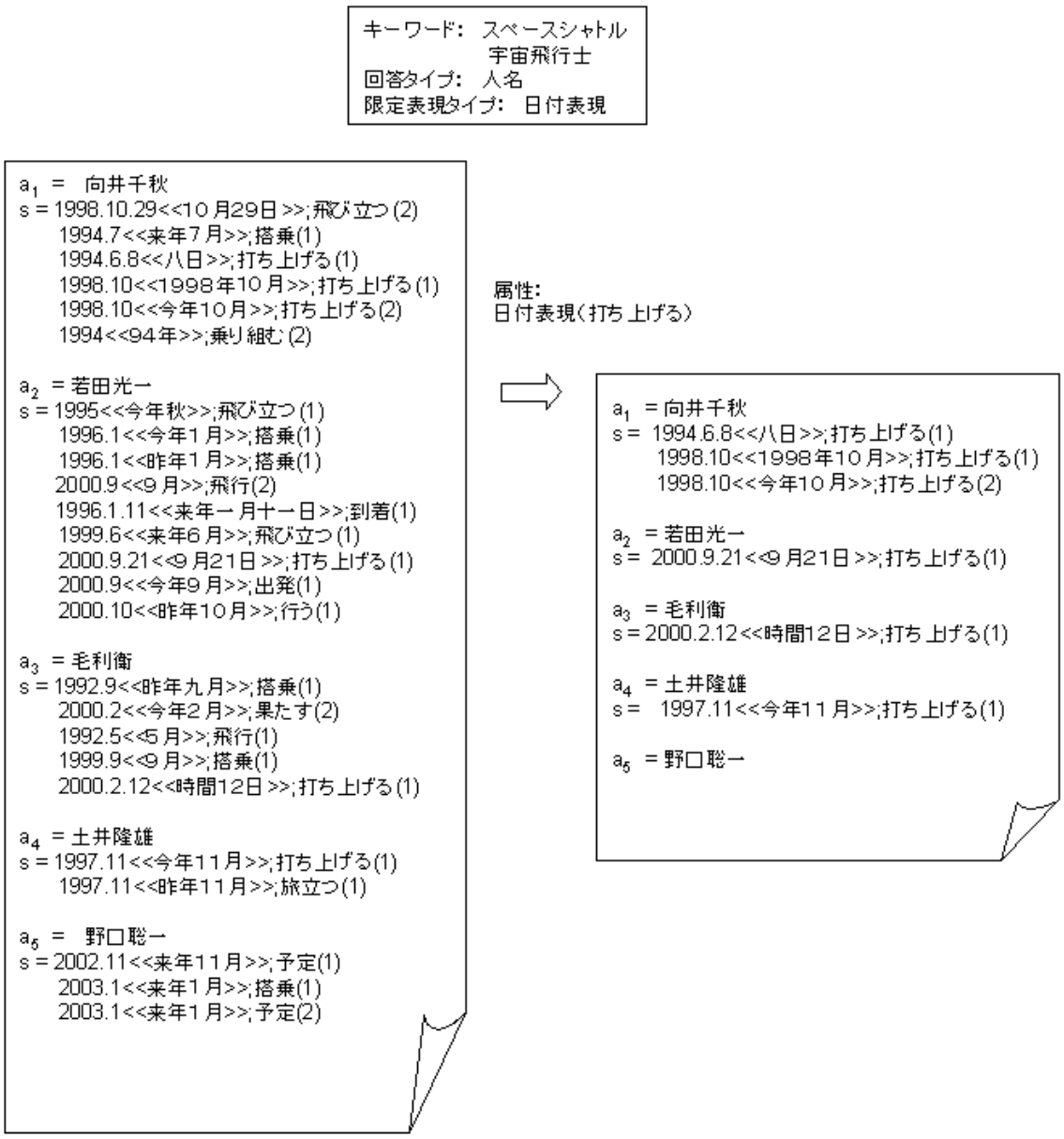

図 5.4: 属性でまとめた限定表現の例3

解答候補数におけるスコア ($\frac{A_{attr}}{A}$)

分母は扱う解答候補数であり, 分子は基準となる限定表現の属性をもつ解答候補数である. このスコアは, 共通の属性を持つ限定表現が多くの解答候補をまたぐことで高くなるスコアである. 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. 表 5.4 の例で挙げると,$S(k_3, s_{succ}, <NUM>+$キロ級) では,A は 3 個であり,A_{attr} は全ての解答候補においてパタンが存在するため 3 個となり $\frac{A_{attr}}{A}$ は 3/3 となる. 一方,$S(k_1, s_{prev}, E3:$キロ級) では A_{attr} が 2 個であるため 2/3 となり, 前者のほうが高いスコアとなる.

限定表現の異なり数によるスコア ($\frac{NT(S(\cdot))}{S(\cdot)}$)

分母は $S(\cdot)$ の要素数, 分子の $NT(S(\cdot))$ は $S(\cdot)$ の異なり数である. この項は, ユーザへの問い返しによって解答候補をどれだけ効率良く絞り込むことができるかを測っている. この項は, 抽出された限定表現が互いに異なる表現であり, かつ共通の属性を持っているということが曖昧性を表していることを意味する. 例として, 表 5.4 における,$S(k_2, se_{prev}, 0046)$ では,S の要素数は「男子」「男子」「女子」の 3 で, 異なり数は「男子」,「女子」の 2 種類となりスコア 2/3 となる.$S(k_3, se_{succ}, <NUM>+$キロ級) では S の要素数は「４８キロ級」,「女子４８キロ級」,「６０キロ級」,「８１キロ級」の 4 で, その異なり数も 4 種類であるのでスコアは 4/4 となる. 前者の場合に問い返しを行うと,「男子」という答えがユーザによって入力された場合に解答が一つに絞り込むことができない. そのため, 後者の方を選択するように考慮したスコアである. この場合,「４８キロ級」と「女子４８キロ級」という単語は数詞以下では同じ文字列であるが, 元の文字列が違うため異なった表現であると判断している.

限定表現の種類におけるスコア ($\frac{\sum_{s \in S(\cdot)} O(s)}{\sum_{attr} \sum_{s \in S(\cdot)} O(s)}$)

分母はグループ内の限定表現総数であり, 分子は属性 $attr$ を持つ限定表現の数である. これは, グループ内で属性を持つ限定表現の割合が大きいほど, その属性の重要性を表しているものとして付けたスコアである. 表 5.4 の $SE(k_3, se_{succ}, <NUM>+$キロ級) で例えると, 要素の数は 2+1+1+1+1=6 であり, その内,「＜NUM＞+キロ級」の属性を持つ要素の総数は 5 個である.

属性のスコア ($P(attr)$)

P(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 5.7 には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.

表 5.7: 属性のスコア

属性	スコア
かっこ	1
数詞＋接尾語	1.3
意味クラス	0.5
末尾 3 文字	1.1
末尾 2 文字	0.6
末尾 1 文字	0.3
日付表現	2

限定表現の平均出現頻度におけるスコア ($\frac{\sum_{s \in S(\cdot)} O(s)}{S(\cdot)}$)

この式の分母は $S(\cdot)$ の種類の数であり, 分子はその $S(\cdot)$ の出現頻度の和である. この項は,S の中の平均出現頻度であり, 出現頻度が大きい限定表現を多く含む S に対して高いスコアがつく. これは, 限定表現の頻度が大きいということは, それだけキーワードを決定づける意味を持っていると考えるためである. 表 5.4 の $SE(k_3, s_{succ}, <NUM>$＋キロ級) を例にとると, 限定表現の種類は 5 種類で, その内「＜NUM＞＋キロ級」の属性を持つものはのべ 6 個であるため, その値は 6/5 となる.

以上の方法によって抽出された限定表現の集合から得られる全ての SE のうち, スコアが最大となるものを選択する. 最大のスコアが得られたグループでは, キーワードとその抽出パタンが示されており, これは, 問い返しを行うために扱うキーワードと抽出パタンも一緒に得られていることを意味する.

5.5 問い返し

曖昧なキーワードの意味をユーザに訪ねる問い返し文を生成する. 問い返し文生成の一番ナイーブな方法は,「ワールドカップの優勝国はどこですか.」という質問に対して,「ワールドカップはサッカーのですか, それともラグビーですか」といったように, ユーザから聞き出すべき限定表現を列挙する方法である. しかし, 解答候補の数が多い場合, 聞き出すべき限定表現の数も多くなり, 問い合わせが不必要に長くなる. そこで, 限定表現を抽象化するなどの工夫が必要である. 本研究では 5 章で述べた限定表現の属性を利用した問い返しを提案する. 選択された限定表現の属性が「かぎ括弧」であり, 限定表現のタイプが連体修飾なら,「何の＜キーワード＞ですか」のように問い返し, 属性が「数詞＋接尾語」で限定表現のタイプが直前の複合名詞なら,「何＋＜接尾語＞＜キーワード＞ですか」, 属性が「意味クラス」で連体修飾であれば,「何の＜意味クラス＞の＜キーワード＞ですか」といったパタンで問い返しを行う. 図 5.2 の属性「意味クラス」の例では,「どの地名のサファ

リパークですか」となり、同図の属性「数詞＋接尾語」では,「何代目黄門ですか」といった問い返し文が生成され, 自然な問い返し文を生成出来ると考える. 現在, 問い返しは未実装である.

5.6 解答選択

問い返し文に対するユーザの返答を受取り, 曖昧なキーワードの意味を決定し, 解答候補の中から最適な解答を選択してユーザに提示する. また. ユーザの応答によってキーワードの曖昧性が解消されない場合は, 再度ユーザへ問い返しを行う. 例として, 図 5.2 では「黄門を演じた俳優は誰ですか」という質問を想定しており, S(黄門, s_{prev}, $< NUM >$ + 代目) というグループが選択されているので, 問い返しは「何代目黄門ですか」になる. この際, ユーザが「５代目」と答えると,「５代目」の限定表現を持つ「里見浩太郎」という回答が選択される. 現在, 解答選択は未実装である.

第6章　評価実験

6.1　実験の手順

4,5章で述べた手法によって曖昧性検出モジュールを評価する簡単な実験を行った. 質問文が曖昧な質問を33個用意し, それぞれに対し解答候補を5つ抽出した. さらに,$S(k_x, type, attr)$のスコアに基づいて限定表現を一つ選択し, それがユーザへの問い合わせ文の内容として適切かどうか人手で判断した. 質問の一覧を表6.1に示す. プライマリキーワードとは主キーワードを表し, セカンダリキーワードはそれ以外のキーワードを表す. 検索年の「all」とは91年から2001年までの12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. 表6.2には質問の中でどのくらい適切な限定表現が抽出できたのか示す. 表6.2(a)は問い合わせ文を生成するために必要な限定表現を抽出することができなかった質問の数であり,(b)は得られた限定表現が適切でない質問の数,(c)は得られた限定表現が適切であった質問の数を表す.

また, 表6.3には, 適切な限定表現が得られた修飾パタンと解答タイプの数を示す.

6.2　考察

表6.2の(a)の結果より, 適切な限定表現が存在しない割合が約20 %程度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80 %程度であった.

表6.2の結果について以下に例を挙げながら結果の解析と考察を述べる.

限定表現抽出成功例

図6.1, 図6.2には限定表現が上手く抽出できたと判断した例を示す. 図では, キーワード, 限定表現のタイプ, 属性でまとめたグループの上位を出力した結果を示す. 各グループは上からキーワード. 限定表現パタン, スコア付けの基準となった解答候補, 属性, スコアの順に並んでおり, その下は解答候補と限定表現を出力する. 図6.1では「シドニー五輪の柔道の金メダリストは誰ですか」との問いを解析した結果を入力とし, 曖昧であると判断した属性を出力した結果の1位と3位を示す. この場合3位に得られる＜生物学的特徴＞とい

表 6.1: 質問一覧

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	キーワードパタン
世界チャンピオン	ボクシング	all	per	hyponym
金メダリスト	柔道, 五輪	all	per	hyponym
金メダリスト	シドニー, 五輪, 柔道	2000-2002	per	other
金メダリスト	バルセロナ, 五輪, 柔道	92,93	per	other
金メダリスト	アトランタ五輪, 柔道	96-98	per	other
金メダリスト	バルセロナ五輪, 水泳	92-94	per	other
金メダリスト	シドニー五輪, 陸上	2000,2001	per	other
金メダリスト	アトランタ五輪, 陸上	96-98	per	hyponym
金メダル	バルセロナ, 五輪, 陸上	92,93	per	other
金メダル	長野五輪	98	per	other
ベスト	ジーニスト	all	per	other
受賞	ノーベル	all	per	hyponym
受賞	ノーベル, 化学賞	all	per	hyponym
受賞	ノーベル, 平和賞	all	per	hyponym
受賞	アカデミー	all	per	hyponym
宇宙飛行士	スペースシャトル	all	per	hyponym
宇宙飛行士	スペースシャトル, 日本人	all	per	hyponym
容疑者	誘拐事件	all	per	other
容疑者	殺人事件	all	per	hyponym
俳優	大河ドラマ	all	per	hyponym
脚本家	家なき子	all	per	hyponym
俳優	黄門	all	per	other
メジャー	日本人, 大リーガー	all	per	other
発売	ドラゴンクエスト	all	time	other
発売	ファイナルファンタジー, ＦＦ	all	time	other
優勝	日本シリーズ	all	org	hyponym
優勝	高校野球	all	org	hyponym
発売	ノートパソコン	99	org	other
優勝国	ワールドカップ	all	loc	hyponym
開催	サミット	all	loc	hyponym
開催地	五輪	all	loc	hyponym
開催	ワールドカップ	all	loc	hyponym
開催地	冬季五輪	all	na	hyponym

表 6.2: 実験結果 1

(a) 適切な限定表現が存在しない	7
(b) 選択した限定表現が不適切	5
(c) 選択した限定表現が適切	21

表 6.3: 実験結果 2

	per	loc	na	org	time	tolal
連体修飾	1					1
直前	2	2	1			5
直後	10				2	12
デ格						0
時間表現	2	1				3
total	15	1	1	2	2	21

う属性よりも 1 位で得られた「＜数詞＞＋キロ級」という表現の方がより適切であると判断した. また, 図には載せていないが 2 位は末尾 3 文字のパタンで抽出された. どちらにしても問い返しは「何キロ級ですか」となる. 図 6.2 では「大河ドラマに出演した俳優は誰ですか」という質問の解析結果を入力した.1 位ではかぎ括弧のパタンが選ばれ,2 位では上位概念のパタンが得られた. この場合, かぎ括弧によるパタンで限定表現が抽出され, 大河ドラマのタイトルについての曖昧性が検出できた.

限定表現抽出失敗例

図 6.3, 図 6.4, 図 6.5 では限定表現の抽出に失敗した例を示す. それぞれの図の上の図は 1 位の出力を属性でまとめる前の限定表現の集合を表す. 下の図はそれを属性でまとめた結果得られた出力である. 図 6.3 は表 6.2 でいう (a) にあてはまる限定表現抽出失敗例であり, 質問に対して得られた限定表現を示す. 図 6.4, 図 6.5 は (b) にあてはまる失敗例であり, 限定表現をグループ化しただけの情報とその結果選ばれた属性を示す.

図 6.3 では「「家なき子」の脚本家は誰ですか」という問いを想定した. この質問では「家なき子」と「家なき子２」というドラマでそれぞれ脚本家が違っているという曖昧性が存在するが,「家なき子」という限定表現から 1 作目であるという情報が得られず, また,5.1 節で述べた限定表現の抽出のルールである, キーワードと同じ限定表現は削除するという制約によって解答候補 1 位「野島伸司」に係っていた限定表現自体が削除されてしまい, 限定表現が抽出されなかった.

図 6.4 では「殺人事件の容疑者は誰ですか」という質問を想定した. この場合, 事件の特徴を表す限定表現は抽出されているが, これらのパタンをまとめる属性が存在せず, 意図し

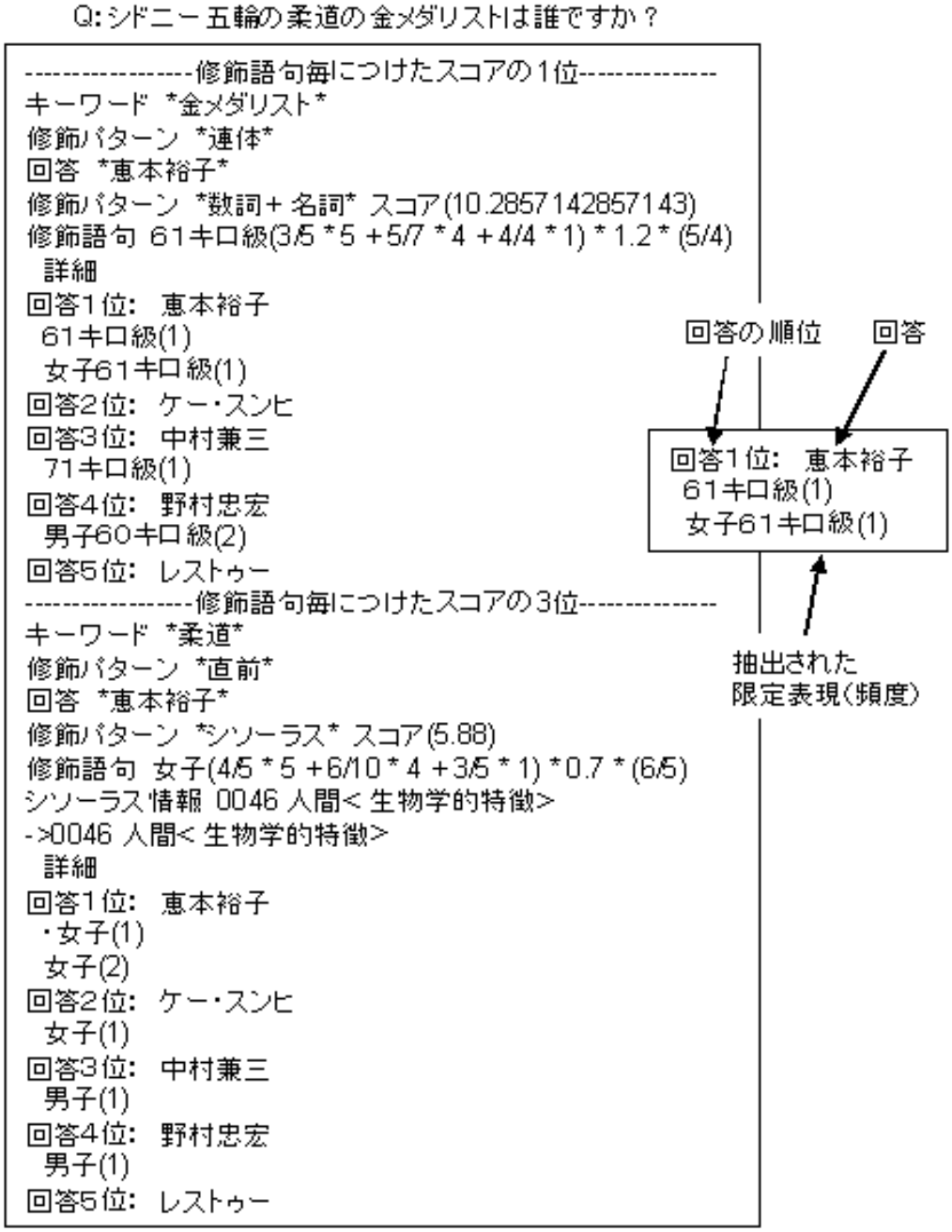

図 6.1: 限定表現抽出例（成功 1）

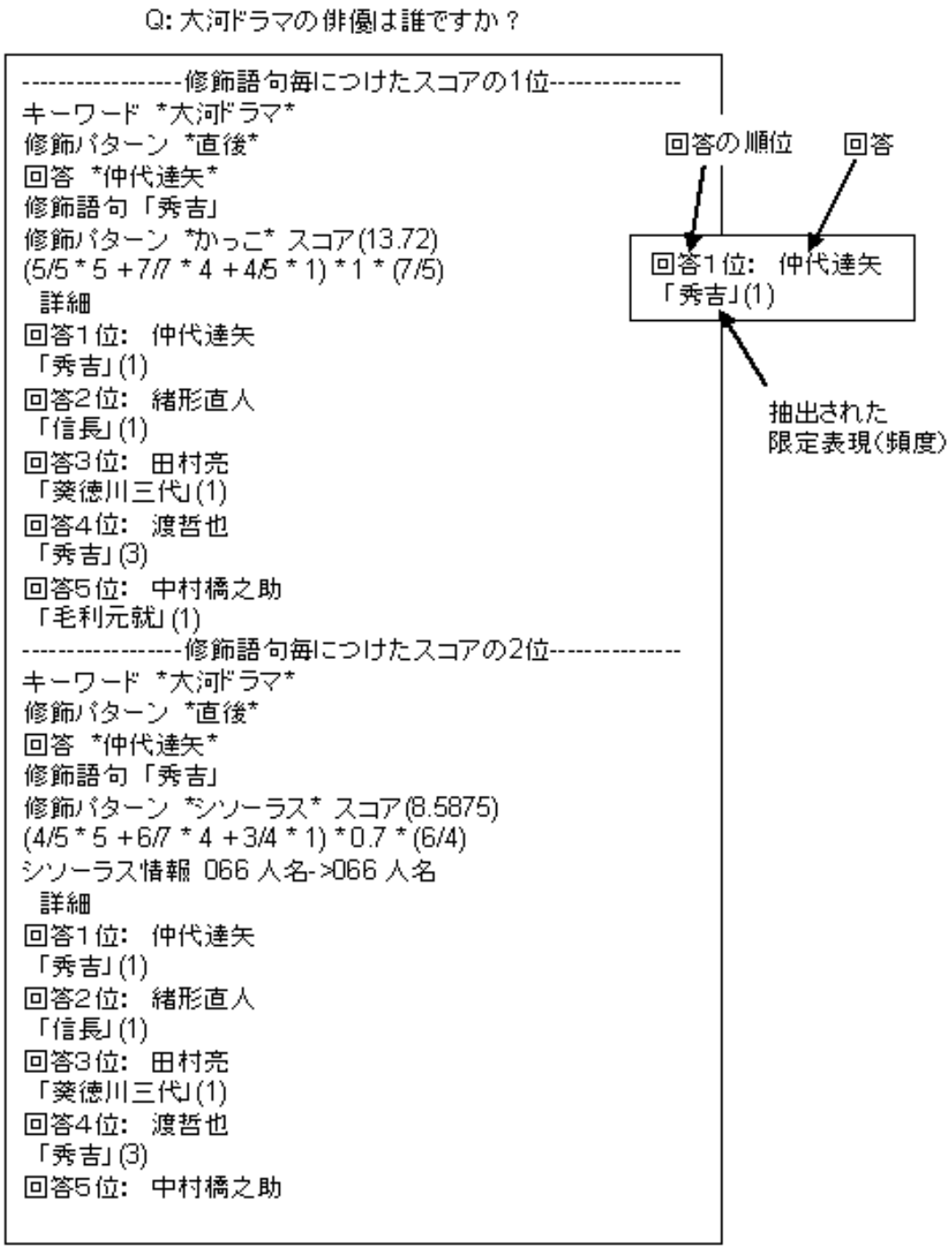

図 6.2: 限定表現抽出例（成功 2）

ない属性が選択された.

図 6.5 では,「長野五輪で金メダルをとった人は誰ですか」という質問を想定した. この場合には,「スピードスケート」や「モーグル」という限定表現は抽出されているがシソーラスに「モーグル」という単語の上位概念がないために限定表現の抽出に失敗した.

図 6.6 では,「アトランタ五輪の陸上の金メダリストは誰ですか」という質問を想定した. シソーラスによって陸上の種目毎の情報は抽出できているが, 問い返しの際には「どのような陸上＜行為＞ですか」のような文になってしまい, 自然な問い返しとはいいがたい. 上位語の抽出に関してもまだ改良の余地が残っていると考える.

表 6.3 では, 各解答タイプに対してどのような修飾タイプで限定表現が抽出されたのかを示す. 質問は適切な限定表現を取れた 21 個の質問を使用した. 人名を問う質問では圧倒的に直後のパタンで限定表現が抽出されることが多いが, それ以外のタイプでは, 直前, 直後のパタンと時間表現がほとんど占めていることが分かった.

実験の結果より, 曖昧な質問に対しての本手法の有用であることの見込みは得られた. しかし, 今回の実験は提案手法の正当な評価ではない. まず, 質問数が 33 と少ないという点がある. また, 今回の実験で用いた質問は, 提案手法を考案する際に検討した質問と同じという意味でクローズドである. さらに, 実験では曖昧な質問のみを入力としていたが, 実際には曖昧でない質問が入力される場合もある. したがって, 曖昧な質問とそうでない質問の両方が入力されるという条件での実験の方がより現実的である.

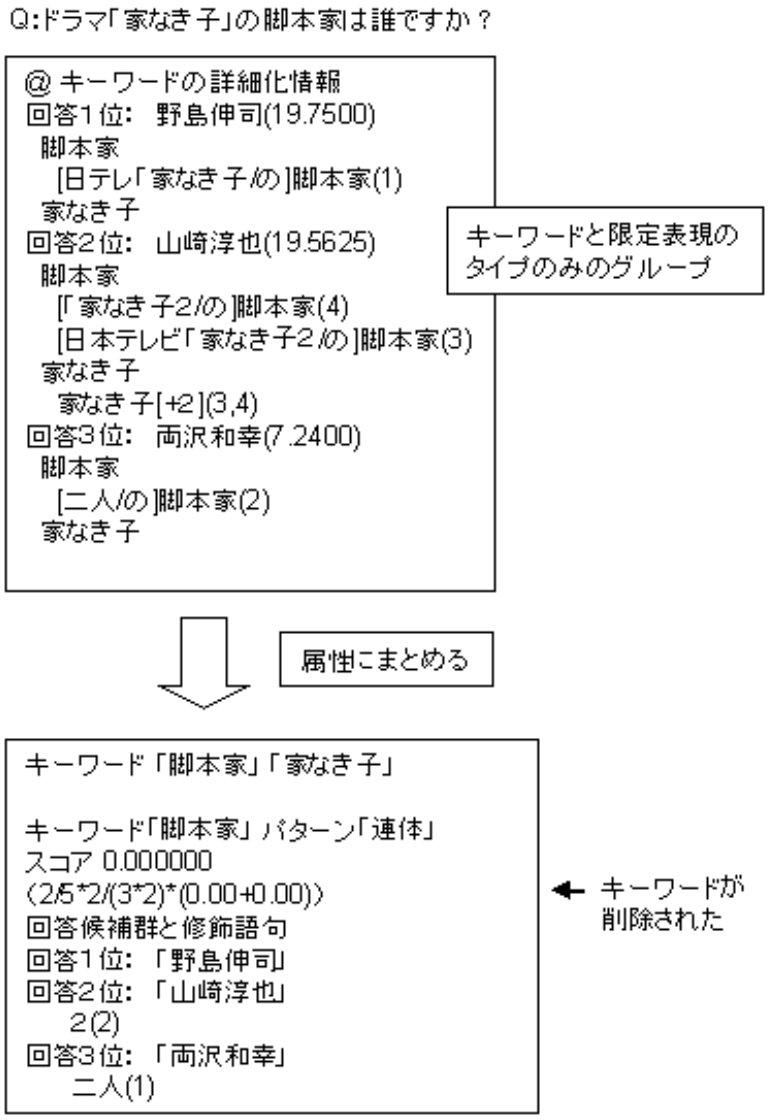

図 6.3: 限定表現抽出例（失敗 1）

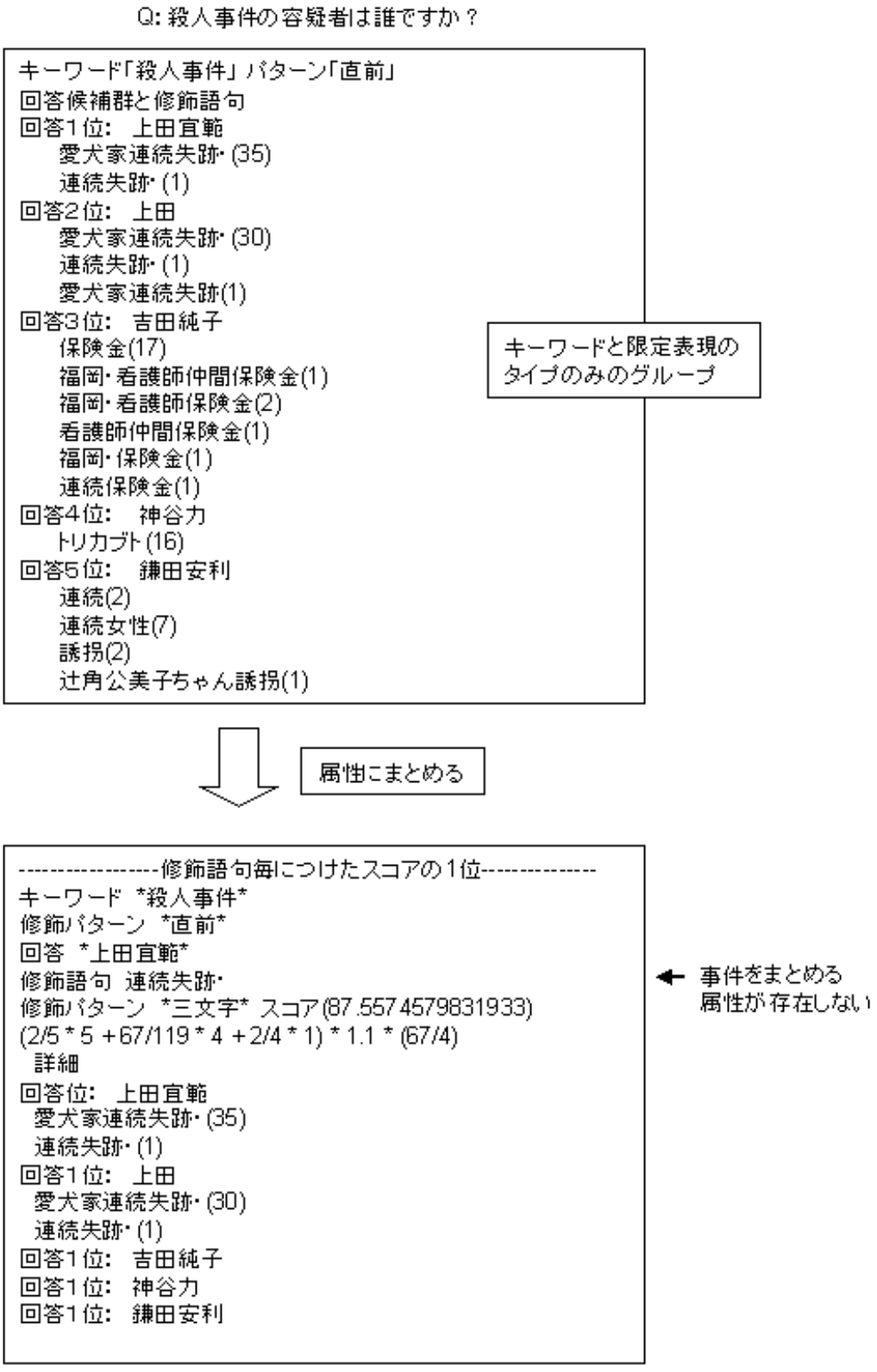

図 6.4: 限定表現抽出例（失敗 2）

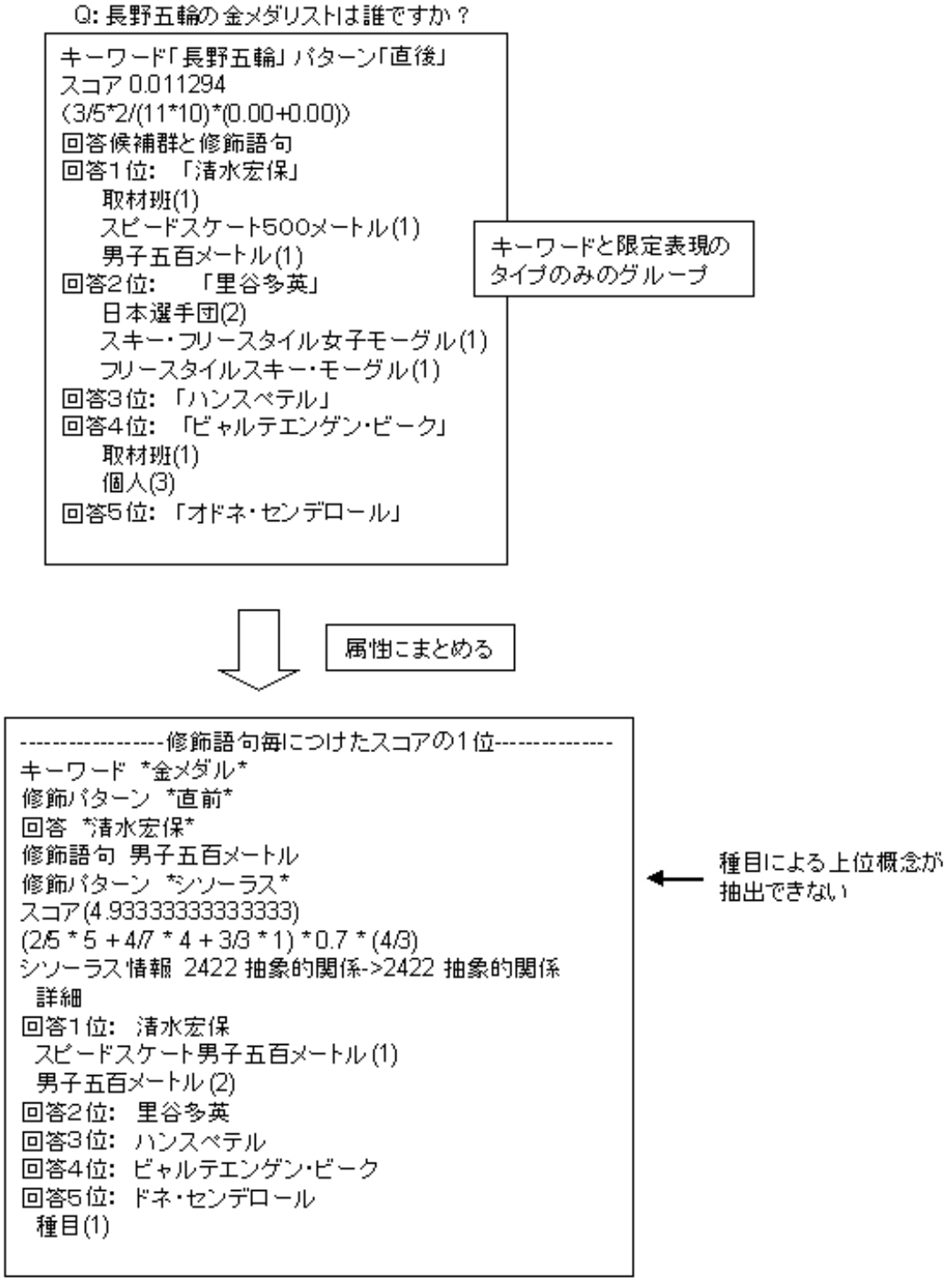

図 6.5: 限定表現抽出例（失敗 3）

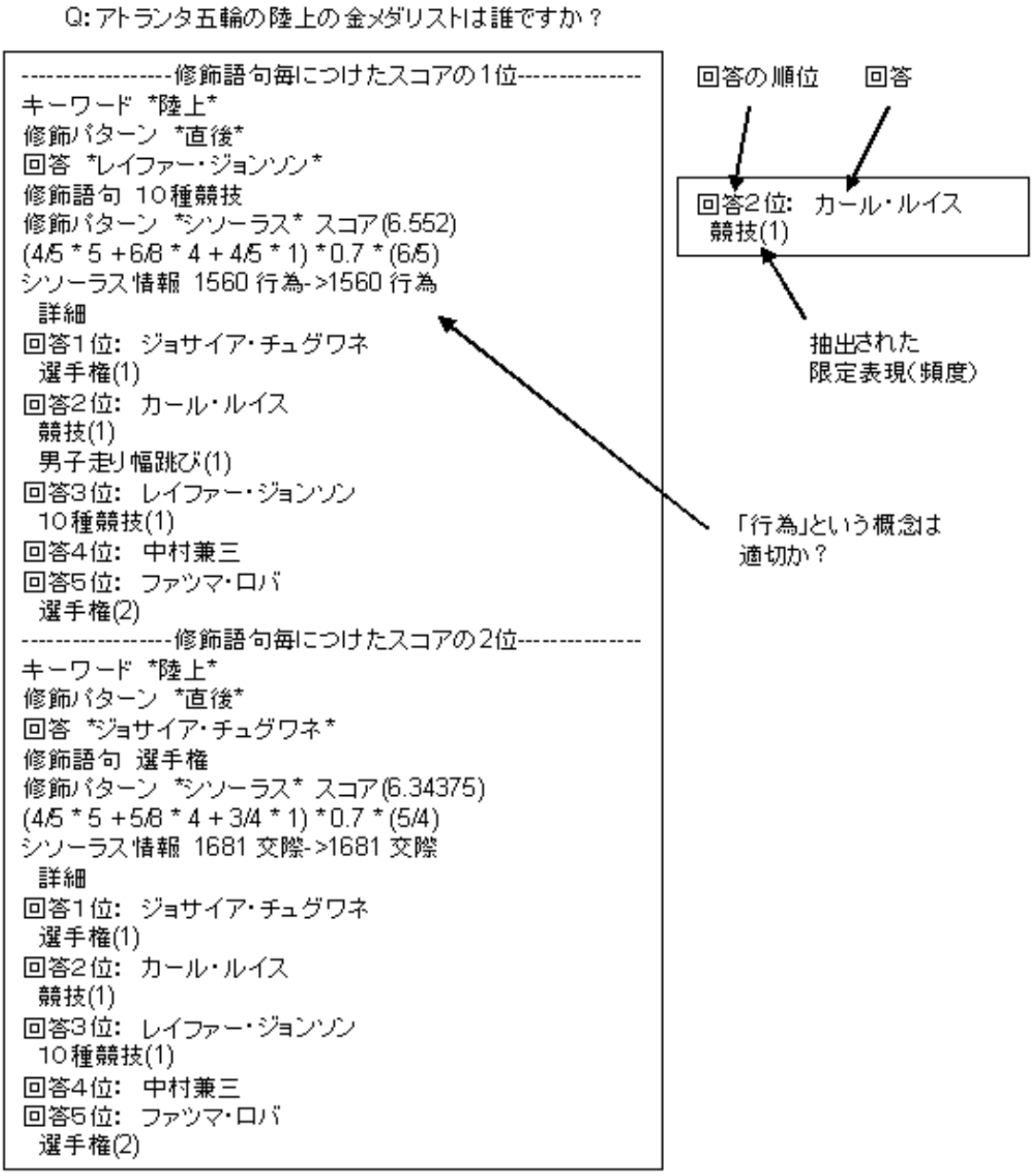

図 6.6: S は抽出されているが判断が難しい例

第7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, 回答候補から曖昧性を検出し, ユーザとの対話を通して適切な解答を選択する対話型質問応答システムについて述べた. また, ユーザの質問文に含まれる曖昧なキーワードを検出し, 問い返し文の内容を決める一手法を提案した.

最後に本研究の今後の課題について述べる.

質問応答システム全般の課題となる問題として, サブシステムの精度の向上がある. 例えば, 質問文の解析に失敗すると, どんなに精度の良い解答抽出システムがあったとしても, その精度に影響が出てしまう. 本研究で主題になっている曖昧性検出のパートは質問文解析, 文書検索, 解答候補抽出の処理を経てから行われるものであり, 解答候補が正しく得られている前提でないと, 提案手法も有効に働かない. そこで, システムを改良する案を以下に述べる.

質問解析

- 扱う解答タイプの増加

  現在扱える質問の種類は 4.1.2 項で述べたように 5 種類の固有表現タグとその他となっている. しかし,5 種類の固有表現タグでは抽出できない解答も存在する. 例として「夏目漱石の名作は何ですか」のような質問の場合では, 解答タイプの設定ができないため, 今後より多くの解答タイプを扱えるようにすることで対応できる質問の幅が広がると考える.

文書検索

- 名詞以外による文書抽出

  本論文では入力キーワードの全文検索によって解答候補抽出の文書群を抽出した. しかし, この手法のみでは, 動詞などの活用する品詞への対応が困難である. 例として「動く」という動詞をキーワードとした場合, 文書では「動いた」のような同じ動詞の活用形でも, 表層の文字列が異なるので抽出できない. 転置インデックスを用いて, 活用する品詞の単語の基本型で検索できるようにすることで, 文書検索の際に, 動詞なども取り扱うことができる.

解答候補抽出

- キーワードパタンの充実

  本論文において解答候補として選択する単語は, キーワードに直接係る単語か, 主キーワードに最も近い名詞の 2 種類であった. そのため, 解答候補として抽出できる単語にかなりの制限があった. 今後はキーワードパタンの充実とキーワードから離れた位置に存在する解答候補の抽出方法を考案することが重要である.

- 扱う解答候補数の決定

  本研究では解答候補数を 5 に設定したが, 実際には質問によって最適な解答候補数が変わる. 解答スコアの調整などにより, 抽出する解答候補数を適切にコントロールすることが望まれる.

曖昧性検出

- 限定表現抽出パタンの充実

  曖昧性を抽出するためには限定表現を抽出することが重要であるため, より多くの限定表現の抽出パタンを考案する必要がある.

- 属性の追加

  6 章でも触れたが, 曖昧性を検出できなかった例として適切な限定表現のグループを属性でまとめることができなかったということが挙げられるため, 属性の種類を増やすことが重要である.

- 曖昧性の有無の検出

  本論文では質問は全て曖昧性が存在するものとして扱った. しかし, 現実問題として, 曖昧性が存在しない質問も数多く存在する. また, 曖昧性を検出することでかえって冗長になってしまうこともある. 例として,「日本の首相は誰ですか」のような質問においては, 普通は現首相を尋ねることが多い. この際に,「いつの首相ですか」などの質問が適切であるかどうかは疑問である. このように曖昧性が存在するが, 問い返しを行わないケースについての対応も必要である.

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] 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.

[2] 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.

[3] Junichi Fukumoto, Tetsuya Endo, Tatsuhiro Niwa. RitsQA: Ritsumeikan question answering system used for QAC-1 *Prpceedings of the Third NTCIR Workshop*, 2002.

[4] Daisuke Kawamura, Nobuhiro Kaji, Sadao Kurohashi. Question and Answering System based on Predicate-Argument Mathing *Prpceedings of the Third NTCIR Workshop*, 2002.

[5] 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.

[6] 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.

[7] Sadao Kurohashi and Wataru Higasa. Dealogue help-system based on flexible matching of user query with natural language knowledge base. *In Proceedings of the 1st ACL SIGdial Workshop on Discourse and Dialogue*, pp. 141-149, 2000.

[8] 前田英作, 磯崎秀樹, 佐々木裕, 賀沢秀人, 平尾努, 鈴木潤. 質問応答システム:SAIQA-何でも答える物知り博士- NTT R&D, Vol.52 No.2, 2003.2.

[9] 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.

[10] Fumito Masui, Masayuki Miyaguchi. MAIMAI:A Question Answering System at NTCIR3 QAC-1 *Prpceedings of the Third NTCIR Workshop*, 2002.

[11] 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.

[12] National Institute of Informatics. *Proceedings of the Third NTCIR Workshop*, 2002.

[13] Masako Nomoto, Mitsuhiro Sato, Hroyuki Suzuki. NTCIR-3 QAC Experimentts at Matsusita *Prpceedings of the Third NTCIR Workshop*, 2002.

[14] Ruck Thawonmas, Takayuki Tomoike, Tomohiro Kawachi, Akio Sakamoto. Exploitation of Newspaper-article Characteristics for Article Retreival and Answer Extraction in QAC Task 2 *Prpceedings of the Third NTCIR Workshop*, 2002.

[15] Satoshi Sekine, Kiyoshi Sudo,, Yusuke Shinyama. NYU/CRL QA system, QAC qusstion analysis and CRL QA data *Prpceedings of the Third NTCIR Workshop*, 2002.

[16] Seungwoo Lee, Gary Geunbae Lee. SiteQ/J:A question Answering System for Japanese *Prpceedings of the Third NTCIR Workshop*, 2002.

[17] Sharon Small et al. HITIQA: Scenario based question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering*, pp. 52-59, 2004.

[18] Sharon Small et al. HITIQA: Towards analytical question answering. In *Proceedings of the COLING, pp. 1291-1297, 2004.*

[19] Toru Takaki, Yoshio Eriguchi. NTT DATA Question-Answering Experiment at the NTCIR-3 QAC *Prpceedings of the Third NTCIR Workshop*, 2002.

[20] Wilensky R., Arens Y., Chin D. Talking to UNIX in English: An Overview of UC. *Communications of hte ACM*, 27(6), pp. 574-593.

[21] 吉浦裕, 片山恭紀, 中西邦夫, 平沢宏太郎. 日本語質問応答システムにおける質問のあいまい性を解消する意味解析方式情報処理学会, Vol.27, No.3, pp.321-329, 1986 Mar.