JAIST Repository
https://dspace.jaist.ac.jp/

Title	ウェブを知識源としたユーザの曖昧な質問に対する質問応答
Author(s)	長内，亘
Citation	
Issue Date	2008-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/4304
Rights	
Description	Supervisor:白井 清昭 准教授，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

ウェブを知識源とした
ユーザの曖昧な質問に対する質問応答

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

長内 亘

2008 年 3 月

修 士 論 文

ウェブを知識源とした
ユーザの曖昧な質問に対する質問応答

指導教官　白井 清昭 准教授

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

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

0610019 長内 亘

提出年月: 2008 年 2 月

Copyright © 2008 by Wataru Osanai

概 要

本論文はウェブ文書を知識源とした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，その曖昧性を検出し，複数の解答をリストとして提示する．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を指す．先行研究が解答を得るための知識源として新聞記事を用いていたのに対し，本論文では知識源としてウェブを用いる．また，解答リストを抽出する手法として，ウェブページにおける表に着目する．本研究では，ユーザに提示する解答リストとなりうる表を発見する手法を提案し，従来のテキスト解析に基づく手法と併用する方法を提案する．

曖昧な質問に対して解答リストを提示するシステムの処理の流れは以下の通りである．まず，ユーザの質問文を解析して，キーワード，解答タイプ，キーワードタイプを抽出する．次に，キーワードをクエリとしてウェブページを検索する．検索されたウェブページから解答リストとなる表を抽出し，表の抽出に成功すればそれをユーザに提示する．表の抽出に失敗した場合は，従来のテキスト解析に基いて解答群を生成する手法を用い，生成された解答群をユーザに提示する．

本手法の評価を行なうために，曖昧な質問 30 個に対して，まずウェブページから表を抽出し，それに失敗した場合は解答群を生成するという方式で解答リストを出力する実験を行なった．その結果，56%の質問に対して，ウェブページから抽出した複数の表の中のいずれかか，テキスト解析によって生成されかつスコアが最大の解答群が正しい解答リストであった．また，83%の質問に対して，表の中に正解があるか，スコアの 10 位以内の解答群の中に正解が含まれていた．これらのケースでは，30 問のうち 9 問についてはウェブページから表を抽出し，残りの質問についてはテキスト解析によって生成された解答群を出力した．また，2 つの手法を組み合わせることで正解が得られる質問の数は増えた．このことから，解答リストを得るために 2 つの手法を併用する提案手法は有効であることがわかった．

目 次

第 1 章　序論 **1**
1.1　研究の背景と目的 1
1.2　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 4
2.3　対話型質問応答システム 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　処理の流れ 7
3.2.1　質問文解析 8
3.2.2　文書検索 11
3.2.3　解答の提示 11

第 4 章　解答リストとなる表の抽出 **12**

第 5 章　テキスト解析による解答群の作成 **17**
5.1　解答候補抽出 17
5.1.1　セグメント分割 17
5.1.2　セグメント検索 18
5.1.3　解答候補抽出 18
5.2　解答群の作成 22
5.2.1　限定表現候補の抽出 22
5.2.2　属性の付与 23
5.2.3　解答群の作成 24
5.2.4　解答群に対するスコア付け 25
5.3　本手法と坂本の手法との比較 27

第 6 章　評価実験 **30**
6.1　実験方法 30
6.2　表の抽出の評価と考察 30
6.2.1　表の抽出の成功例 31
6.2.2　表の抽出の失敗例 32
6.2.3　誤って抽出した表の例 33
6.3　テキスト解析による解答群の作成の評価と考察 35
6.3.1　解答群作成の成功例 37
6.3.2　解答群作成の失敗例 37
6.4　組み合わせ手法の評価と考察 38

第 7 章　参照日付表現の処理 **39**
7.1　提案手法 39
7.2　評価 41

第 8 章　結論 **43**

付 録 A　実験に用いた質問一覧 **47**

図 目 次

3.1 提案システムの出力例 (解答群) 7
3.2 解答群に相当するウェブページの表の例 8
3.3 本システムの処理の流れ 9

4.1 セルが結合された表の例 1 13
4.2 セルが結合された表の例 2 14
4.3 解答の存在のチェック 15

5.1 検索されるセグメントの例 18

6.1 表の抽出の成功例 32
6.2 表の抽出の失敗例 1 33
6.3 表の抽出の失敗例 2 33
6.4 表の抽出の失敗例 3 34
6.5 表の抽出の失敗例 4 34
6.6 誤って抽出した表の例 1 35
6.7 誤って抽出した表の例 2 36
6.8 解答群作成の成功例 37
6.9 解答群作成の失敗例 38

7.1 日付表現の抽出の例 41

表 目 次

3.1 解答タイプ 10
3.2 キーワードタイプ 10

4.1 解答タイプと固有表現タグの対応表 15

5.1 セグメント分割に用いたタグ 18
5.2 解答候補が満たすべき条件 19
5.3 解答候補抽出パターンとスコア 20
5.4 品詞情報のスコア 21
5.5 属性のスコア 26
5.6 限定表現の抽出パターンのスコア 27
5.7 属性のスコアの比較 29
5.8 限定表現の抽出パターンのスコアの比較 29

6.1 質問の例 30
6.2 質問解析の結果 31
6.3 表の抽出の実験結果 31
6.4 解答群作成の実験結果 36
6.5 解答群作成の実験結果 38

7.1 更新を示唆する表現 40
7.2 日付表現抽出の実験結果 42

A.1 予備実験の質問一覧 47
A.2 予備実験の質問設定一覧 49
A.3 評価実験の質問一覧 50
A.4 評価実験の質問設定一覧 52

第1章　序論

1.1　研究の背景と目的

質問応答システムはユーザの質問に対する解答をテキストの中から探して答えるシステムである．ところが，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つに絞ることが出来ない場合がある．曖昧な質問の例をあげると「アカデミー賞を受賞したのは誰ですか？」といった質問の場合，主演男優賞のラッセル・クロウ，主演女優賞のジュリア・ロバーツ，監督賞のスティーブン・ソダーバーグ等が解答候補として考えられ,「アカデミー賞」に部門の曖昧性がある為に解答を1つに絞ることが出来ない．このような問題を考慮し，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に対応した質問応答システムの研究がある．松本 [1] や坂本 [2] は質問の曖昧性を検出する質問応答システムを提案している．

松本は「アカデミー賞を受賞したのは誰ですか？」という質問に対し，以下のような解答リストをユーザに提示する手法を提案した．

主演男優賞：　ラッセル・クロウ
主演女優賞：　ジュリア・ロバーツ
監督賞：　スティーブン・ソダーバーグ

一方，坂本はシステムからユーザに問い返すことで適切な解答を選択する手法を提案した．

システム：　「アカデミー賞の何賞ですか？」
ユーザ：　「主演男優賞です」
システム：　「受賞者はラッセル・クロウです」

これらの研究では知識源となるテキスト集合として新聞記事を用いていた．しかし，新聞記事から得られる情報には限りがある．一方，ウェブ上には多くの情報が存在する．したがって，解答を探すテキストとしては新聞記事よりもウェブのほうが望ましいと考えられる．またウェブを対象とした質問応答システムもいくつかあるが，質問の曖昧性は考慮されていない．そこで本研究では，ウェブを知識源とし，ユーザの曖昧な質問を受け付けることのできる質問応答システムを提案する．

本研究では，松本，坂本らの手法を基盤とし，ウェブから適切な解答を得るために必要な改変に取り組む．これらの先行研究は曖昧な質問に対して知識源となるテキスト中の文を解析し，解答リストを動的に生成していた．これに加え，本研究ではウェブページにおける表に着目する．ウェブを知識源とした質問応答の場合，質問によってはウェブページ中の表の中にその質問に対する解答リストが存在する場合がある．本論文では，ユーザに提示する解答リストとなりうる表を発見する手法を提案し，従来のテキスト解析に基づく手法と併用する方法を提案する．

1.2 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，本研究で提案する質問応答システムの概要について述べる．
4 章では，解答リストとなる表を抽出するための手法を述べる．
5 章では，テキスト解析によって解答群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，参照日付表現を適切に取り扱うために，テキストが書かれた日付を同定する手法について検討する．
8 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

質問応答システムとは，自然言語を入力とし，ユーザの質問に対する解答をテキストの中から探して答えるシステムである．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．

1. 質問文の解析
2. 文書検索
3. 解答抽出

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，現在の質問応答システムに関する研究は，ユーザの質問に対して複数の解答が得られる場合でも，その中のうち 1 つが正し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を研究の対象とする場合が主流である．

賀沢らは TREC-10 における質問応答システムを構築した [3]．このシステムでは SVM による固有表現の認識とヒューリスティックによる同格関係の発見を行なっている．まず，人手で作成した規則によって固有表現の候補を抽出する．その後に SVM を用いて PERSON，ORANIZATION，LOCATION，OTHER の 4 つのクラスに分類を行なう．同格関係の発見については，賀沢らはコンマの役割について着目し，コンマの役割を以下のように分類した．

1. 構文的移動
   “When I was a kid, things were simple.”
2. 等位関係
   “cats, dogs and birds”
3. 同格関係
   “George, the son of the former president, is a popular man.”

そして以下のようなヒューリスティックに基づき同格関係を抽出した．

1. 文章が従属接続詞から始まるならば，左端のコンマは構文的移動を意味する．

2. 文章が “名詞句， 名詞句 等位接続詞 名詞句” と続いているならば，これらのコンマは等位を意味する．

3. 1，2 に該当しないコンマは同格関係を表わすとみなす．

2.2 リスト型質問応答システム

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群を見つけ，解答の信頼性を表わすスコアを付与し，上位のスコアの解答を出力する．したがって解答が複数得られた場合でも，最上位のスコアの解答を正しい解答として出力する．例えば「世界三大珍味はなんですか」という質問に対して従来のシステムを用いると,「キャビア」「フォアグラ」「トリュフ」のいずれかを出力する．しかし，これらはいずれも世界三台珍味の 1 つで正解である．ユーザが解答を求める意図は,「三大珍味」とあるように 3 つの解答全てを知りたいことであると考えられるため，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，複数の解答を返すべき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．

石下らは複数の解答をもつ質問に対応する為に，解答候補の集合のスコア分布が正解集合の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 アルゴリズムを用いて分離して，正解側の分布に由来すると推定できる解答候補を正解として出力している [4]．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解答候補群のスコアを再度計算し，解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，不正解の解答も正解集合に加えてしまうという傾向がみられた．

福本らは質問応答システムによって抽出されたスコア上位 20 件の解答候補を用いて解答リストを作成した [5]．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，以下の方法によってスコア付けを行っている．

- 同じ文の中にある解答候補のスコアの重み
- キーワードと最も近い位置にある解答候補の重み
- 上記 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 animals」は解答数が 3,「combi」は解答数が 2,「who and who」は解答数が 2 であることを示す．

松本らは曖昧な質問に対して解答群をリスト表示するシステムの構築を行った．例えば「アカデミー賞の受賞者は誰ですか」という質問の場合，アカデミー賞には主演男優賞，主演女優賞，監督賞など様々な賞があり，それぞれの「賞」に対して解答が異なっている．そこで，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，解答と限定表現の組を作り，それをリストとしてユーザに提示した．先ほどの例は「主演男優賞：ラッセル・クロウ」「主演女優賞：ジュリア・ロバーツ」「監督賞：スティーブン・ソダーバーグ」のようなリストになる (：の前は限定表現，後は解答を表わす)．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．

2.3 対話型質問応答システム

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．

黒橋らは京都大学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 [6]．まずユーザの質問を構文解析し，発話タイプに分類する．次に，ユーザの質問を解析した情報と知識ベースのマッチングを行い，最も類似度の高い部分を見つけて，それに対応する解答をユーザに提示する．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．このシステムは事実を問う What 型，方法を問う How 型，症状を提示しその対処を求める Sympton 型の質問に回答し，さらに以前の質問に対して返答する Answer 型および以前の質問の修正・追加をする Addition 型の発話を文脈に応じて適切に解釈する枠組みをもっている．メディアセンターに対して要求する Request 型については対象外としている．ユーザとの対話には，未知語の問い返し，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問い返し，挨拶に対しての返事を行なうことができる．

清田らは Windows 環境の利用者を対象とした対話的質問応答システム「ダイアログナビ」を構築した [7]．マイクロソフトが既に一般に公開しているテキスト知識ベースをデータベースとして利用し，ユーザへの問い返しは「対話カード」に従って行なう．対話カードとは，あらかじめ典型的な質問に対して，どのような問い返しを行な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を調べる．一致する場合は，対話カードに従い問い返しを行なう．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のうち，ユーザの質問と一致しなかった箇所のことである．例えば,「表を作成する」という質問に対し，テキスト知識ベースに「PowerPoint で表を作成する」という文と「Word で表を作成する」という文が抽出された場合,「PowerPoint で」と「Word で」の部分を抽出し，ユーザに提示して解答を選択させる手法である．

坂本らは，松本らと同様の手法で曖昧性を検出し，曖昧なキーワードの意味をユーザに問い返すことで適切な解答を得る手法を提案している．ユーザとの対話には，問い返し用の 3 種類テンプレートを用いて行なう．1 つ目は，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，例えば,「何賞」というように，解答の曖昧性全体を表わす語「賞」に疑問詞「何」をつけた表現である.「アカデミー賞の受賞者は誰ですか」という質問の場合は，アカデミー賞には主演男優賞，主演女優賞，監督賞など「賞」という観点で曖昧性があるので，解答の曖昧性全体を表わす語は「賞」となる．3つ目は，問い返し主題を含まない問い返し文を生成するテンプレートである．以上の3つのテンプレートから複数の問い返し文を生成し，n-gram の頻度と Web 検索エンジンのヒット数によるスコア付けに基づき最適なものを1つ選択した.

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．質問が曖昧であるというのは，質問文が明確ではなく，キーワードの意味が一意に決められず，解答を絞り込めないことを意味する．例えば,「ノーベル賞の受賞者は誰ですか」という質問の場合，ノーベル賞には平和賞や医学賞といった複数の部門が存在するため，解答が1つに絞り込めない．この場合，ノーベル賞というキーワードには曖昧性があると考え，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ユーザに提示する．これによりユーザが自分の質問の曖昧性に気づ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.

- ウェブを知識源とする
  2.2 節，2.3 節で紹介した曖昧な質問に対応した質問応答システムとの違いとして，知識源となるテキスト集合として新聞記事ではなくウェブを用いるという点が挙げられる．新聞記事から得られる情報の量や分野には限りがあるのに対して，ウェブはテキストの量がはるかに多く，より多様な情報が存在する．したがって，解答を探すテキストとしては新聞記事よりもウェブのほうが望ましいと考えられる.

- 解答リストを含む表を抽出する
  2.2 節，2.3 節で紹介した曖昧な質問に対応質問応答システムとの違いとして，解答リストを含む表をウェブページ中から抽出するという点が挙げられる．曖昧な質問に対して知識源となるテキスト中の文を解析し，解答リストを動的に生成する従来の手法と，解答リストを含む表を抽出し，ユーザに提示する手法を併用することで，より多様な質問に対応できると考えられる.

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った．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，キーワード個々の意味と共に解答を出力する．例えば「アジアカップで優勝した国はどこですか」という質問に答える場合を考える．ここで，アジアカップには「2007 年」,「2004 年」のように年が複数存在するため，解答をひとつに絞ることができない．本研究ではこの「2007 年」「2004 年」といった表現をキーワード (この場合は「アジアカップ」) の意味を限定することから限定表現と呼ぶ．曖昧な質問に対する解答リストの抽出は以下の二段階の手順で行なう．

1. 解答リストを含む表の抽出
   ウェブページの中には，図 3.1 のような解答群に相当する表が存在するときがある．図 3.2 にその例を示す．このような表がウェブにおけるいずれかのページに存在する場合には，それを抽出しユーザに提示する．

2. テキスト解析による動的な解答群の生成
   解答リストを含む表の抽出に失敗した場合は，従来のテキスト解析による動的な解答群の作成を行なう．まず，限定表現とそのキーワードならび解答候補で 3 つ組を作る．図 3.1 のように共通する属性 (この場合は「年」が共通する属性) を持つ限定表現を含む 3 つ組をまとめて解答群を作り，ユーザに提示する．

	解答	限定表現
(1)	イラク	2007 年のアジアカップ
(2)	日本	2004 年のアジアカップ
(3)	サウジアラビア	1996 年のアジアカップ

図 3.1: 提案システムの出力例 (解答群)

3.2　処理の流れ

本システムにおける処理の流れを，図 3.3 に示す．以下にその概要を示す．

開催年	開催国	優勝	準優勝
2004	中国	日本	中国
2000	レバノン	日本	サウジアラビア
1996	UAE	サウジアラビア	UAE
1992	日本	日本	サウジアラビア
1988	カタール	サウジアラビア	韓国

図 3.2: 解答群に相当するウェブページの表の例

- 質問文解析
  ユーザの質問を受け付け，質問文の解析をする．

- 文書検索
  質問文解析で得られたキーワードをクエリとしてウェブページを検索する．

- 解答リストの抽出 (A)
  3.1 節の 1 で述べたように，ウェブページから表を抽出する．

- 解答リストの抽出 (B)
  3.1 節の 2 で述べたように，テキスト解析によって解答群を生成する．

(A) の処理を (B) の処理の前に行なうのは，ウェブページから抽出された表は，解答に対する答えとして適切である可能性が高いからである．その反面，全ての質問に対して解答となる表が存在するとは限らない．そのため，表を抽出できなかったときは，(B) の処理によって動的に解答群を生成する．

図 3.3 に示した処理のうち，従来の手法とほぼ同じ処理を行なう質問文の解析，文書検索，解答の提示については以下の項で述べる．本研究のテーマである解答リストとなる表の抽出は 4 章で，解答群の作成の詳細は 5 章で述べる．

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．

- 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単語である．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．

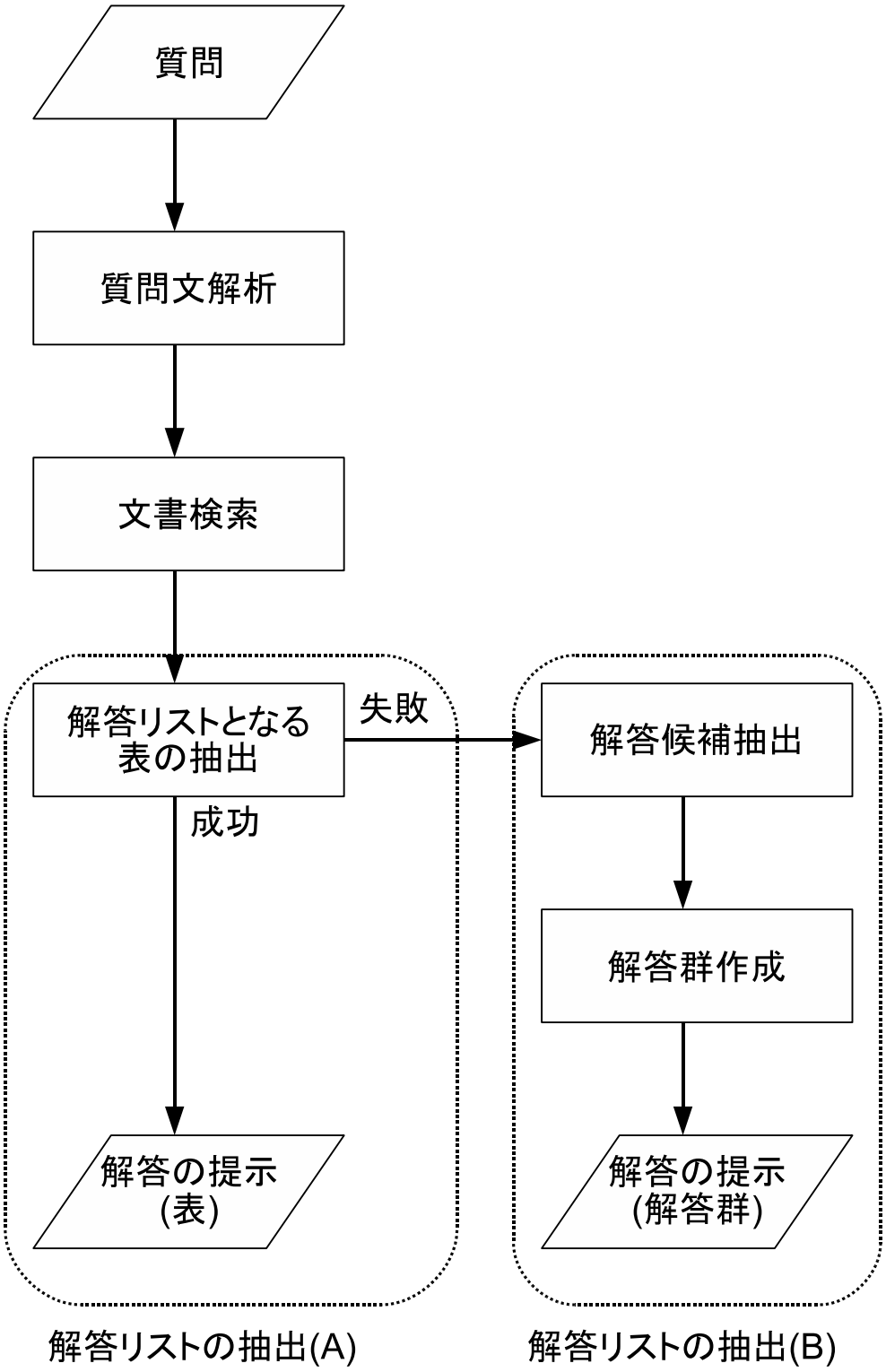

図 3.3: 本システムの処理の流れ

プライマリキーワード

文書検索と解答候補抽出のための中心となるキーワードである．解答と最も関係の深いキーワードであり，質問文中の主題に当たる名詞などが該当する．質問文中のプライマリキーワードは 1 つとする．

セカンダリキーワード

文書検索と解答候補抽出のためのキーワードである．セカンダリキーワードは一般に複数あるものとする．

- 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「数値」,「その他」の 7 種類を設定した．解答タイプの詳細を表 3.1 に示す.

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]	「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
数値 [am]	「いくつ」を尋ねるタイプで数量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 3.2 にキーワードタイプの一覧を示す.

表 3.2: キーワードタイプ

名前	パターン	キーワードと解答候補の関係
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.「日本有線大賞を受賞したのはだれですか」という質問文を解析する．形態素解析の結果から「日本有線大賞」,「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「日本有線大賞」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ので，キーワードのタイプが「agent」と決まる.

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.

3.2.2 文書検索

質問文解析で得られたキーワードをクエリとしてウェブページを検索する．検索エンジンには TSUBAKI[9] を用いた．TSUBAKI は検索のためのインデックスとして単語 (代表表記)，係り受け関係，文字トライグラムを用いることで，表記のゆれやクエリの意味を考慮した検索エンジンであり，質問応答システムの為の検索エンジンとして適しているといえる．本研究では検索結果の上位 100 件のウェブページを取得し，処理の対象とした.

3.2.3 解答の提示

文書検索後の処理の流れは以下の通りである．検索されたウェブページの中に解答リストとしてふさわしい表があれば，それを抽出してユーザに提示する．解答リストとなる表が見つからなかった場合には，ウェブページ内のテキストを解析して複数の解答候補を抽出し，それらをまとめた解答群をユーザに提示する.

第4章　解答リストとなる表の抽出

この章では，ウェブページから解答リストとなる表を抽出する手法を述べる．ここでは，キーワード，解答タイプ，表抽出の対象とするウェブページの集合を入力とし，解答リストとしてふさわしい表を抽出することを目的とする．解答リストとなる表を抽出するための処理の流れを以下に示す．

1. 表候補の検出
2. プライマリキーワードによる表の選定
3. セカンダリキーワードによる表の選定
4. 固有表現タグによる表の選定
5. 解答リストとなる表の提示

これらの処理についての詳細を以下に述べる．

ステップ1：表候補の検出

table タグで定義されている表を検出する．しかし，表が属性と複数の解答を表わすには少なくともセルの数が3つ以上なくてはならず，行と列のセルの数が両方とも2つ以下の場合，その表から複数の解答を得られる可能性は低い．本研究では曖昧性を含む質問に対する解答リストとなる表の抽出を目的としているので，このような表は候補から除外する．また，ウェブページには画像やテキストなどで表が記述されている場合があるが，これらは抽出の対象外とする．

ステップ2：プライマリキーワードによる表の選定

一般に，表の1行目または1列目はなんらかの属性を表わすとみなせる．また，プライマリキーワードは解答リストとなる表の中において解答の属性を表わすことが多いと考えられる．もし，プライマリキーワードと表の属性が一致していれば，その表は解答を含む可能性が高い．そこで表の1行目または1列目にあるセル内のテキストとプライマリキーワードが一致する場合に，その表を解答リストを含む表の候補として抽出する．また，表のセル内の文字列とプライマリキーワードが完全に一致しなくても，以下の両方の条件を満たす場合は，その文字列はプライマリキーワードを表わすとみなす．

条件 1 セル内の文字列が複合名詞である

条件 2 セル内の文字列の末尾，または最後の 1 文字を除く末尾がプライマリキーワードと一致する

例えば,「アカデミー賞を受賞したのはだれですか」というような質問の場合，質問解析によってプライマリキーワード「受賞」が抽出される．しかし，解答リストとなる表の属性を表わす単語として「受賞者」や「今年度受賞者」等も存在する．これらはキーワード「受賞」と完全に一致はしていないが，ほぼ同じ意味を持つと考えられる．上記の条件を用いることによって，キーワードと 1 行目または 1 列目のセル内のテキストが完全に一致していない場合にも解答を含む表を抽出することができる．なお，条件 1 に示す複合名詞とは Juman[12] を用いてセルの文字列を解析し，その文字列が名詞，接尾辞，接頭辞，記号，カタカナ，ナ形容詞で構成されているものを指す．また，キーワードのマッチングを行なう際に，Juman の代表表記を用いることで柔軟なマッチングを行なう．代表表記については後述する．

また，1 行目または 1 列目のセルが結合されているとき，そのセルは表のキャプションやタイトルを意味する場合や，解答の属性のさらに上位の属性を表わす場合であると考えられる．このような場合には次の行または列をプライマリキーワードの検索を行なう対象とした．図 4.1，図 4.2 にセルが結合された表の例を示す．図 4.1 の表において，結合されている 1 行目の「ワールドカップ年表」は表のキャプションの役割を果たしている．一方，図 4.2 の表において，結合されている 1 行目の「順位」は優勝，2 位，3 位の上位の属性を表わしている．

ワールドカップ年表		
開催年	開催国	優勝国
1930 年	ウルグアイ	ウルグアイ
1934 年	イタリア	イタリア
1938 年	フランス	イタリア
1950 年	ブラジル	ウルグアイ
1954 年	スイス	西ドイツ

図 4.1: セルが結合された表の例 1

ステップ 3：セカンダリキーワードによる表の選定

表にプライマリキーワードが出現している場合でも，その表が質問に対する解答群を含んでいるとは限らない．そこで，セカンダリキーワードを参照し，質問のトピックと関連の深い表のみを抽出する．具体的には，全てのセカンダリキーワードが以下の 3 つのいずれかの場所に存在するかを調べ，存在しない場合はその表を候補から除外する．ステップ 2 と同様に，キーワードのマッチングを行なう際には，Juman の代表表記を用いる．

	順位		
	優勝	2 位	3 位
2007 年	巨人	中日	阪神
2006 年	中日	阪神	ヤクルト
2005 年	阪神	中日	横浜
2004 年	中日	ヤクルト	巨人
2003 年	阪神	中日	巨人・ヤクルト

図 4.2: セルが結合された表の例 2

1. ウェブページの title タグの中
   ウェブページの title タグの中にあるキーワードはページ全体のトピックを表わすと考えられる．キーワードが title タグにある場合は，そのページは質問と関連が深いとみなし，ページ内の表を抽出の対象とした．

2. 表のキャプション
   表のキャプションの中にあるキーワードは表全体のトピックを表わすと考えられる．キーワードが表のキャプションにある場合は，その表は質問と関連が深いとみなし，抽出の対象とした．

3. 表の前にある 3 つのセグメント
   表の前のセグメントの中にあるキーワードは表全体のトピックや表のキャプションを表わすと考えられる．キーワードが表の前のセグメントの中にある場合は，その表は質問と関連が深いとみなし，抽出の対象とした．セグメントの詳細については 5.1 節で述べる．

なお，3. の表の前にある 3 つのセグメントという条件は，解答リストとなる表とキーワードの出現場所を予備実験によって調べ，表抽出の再現率が最も高くなる値を設定した．また，表の後ろのセグメントにキーワードが出現する頻度は低く，表のトピックを特定する手がかりにはならなかった．予備実験に用いた質問を付録 A.1 に示す．

ステップ 4：固有表現タグによる表の選定

ステップ 2 で検出したセルが表の 1 列目にあるとき，そのセルと同じ行にあるセルのテキストが解答を含むかを調べる．同様に，ステップ 2 で検出したセルが表の 1 行目にあるときは，そのセルと同じ列にあるセルのテキストが解答を含むかを調べる．セルが 1 行 1 列目の場合は両方を調べる．図 4.3 に解答を調べる手順を示す．(A) は 1 列目にプライマリキーワードが検出された場合の例を表わす，(B) は 1 行目にプライマリキーワードが検出された場合の例を表わす．また，結合されたセルの検出により，2 行目または 2 列目以降のセルとプライマリキーワードが一致している場合は，一致しているセルの場所を開始

位置として同じ行または列を調べる．図 4.1 の表の例で仮にプライマリキーワードが「優勝」の場合，2 行目のセル「優勝国」がステップ 2 で検出される．そして，そのセルを開始位置として同じ列のセルが解答を含むかを調べる．

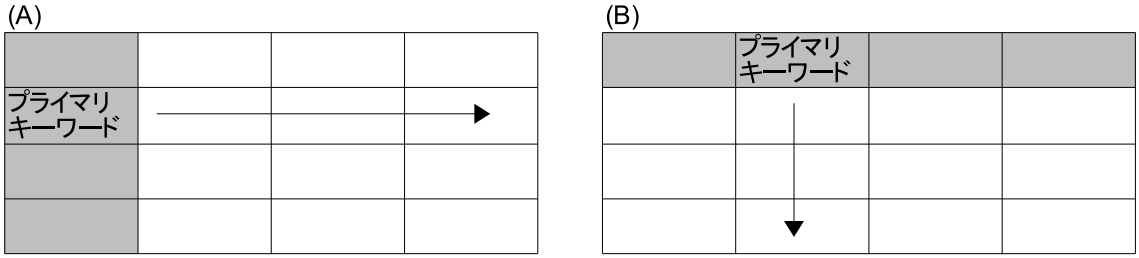

図 4.3: 解答の存在のチェック

表の 1 行または 1 列が解答を含むかどうかの判定は以下の手順で行なう．

1. セル内のテキストが複合名詞であるかを調べる．複合名詞の判定の条件はステップ 2 の条件と同じである．

2. セル内のテキストを南瓜 [11] で解析し，固有表現タグを割り当て，その固有表現タグがユーザの解答タイプと一致しているかを調べる．

3. 1 行または 1 列内のセルのうち，複合名詞であり，固有表現タイプと解答タイプが一致しているセルの割合を調べ，それが 0.3 以上の時にはその表を解答リストを含む表として抽出する．

また，0.3 という閾値は抽出された表の適合率と再現率を付録 A.1 の質問を用いた予備実験によって調べ，バランスが最もよくなる値を設定した．

なお，本研究の手法では南瓜を用いて固有表現タグの抽出ができる人名，国，地名，組織名，時間の解答タイプの場合のみを対象としている．解答タイプが数値，その他の場合は表の抽出は行なわない．表 4.1 に解答タイプと南瓜が出力する固有表現タグの対応を示す．

表 4.1: 解答タイプと固有表現タグの対応表

解答タイプ	固有表現タグ
人名	⟨PERSON⟩
国	⟨NATION⟩
地名	⟨LOCATION⟩
組織名	⟨ORGNIZATION⟩
時間	⟨DATE⟩

代表表記について

本研究ではキーワードによるマッチングを行なう際に，Juman で用いられる代表表記を用いることによってキーワードの表記のゆれに対応し，表抽出の再現率の向上を図る．例えば，果物の林檎はカタカナの「リンゴ」，ひらがなの「りんご」，漢字の「林檎」と表わすことができる．Juman でこれらの解析を行なうと,「代表表記:林檎」という共通する表記を取得できる．キーワードとテキスト中の単語でマッチングを行なう際には，それぞれの代表表記が一致しているかをチェックする．

第5章　テキスト解析による解答群の作成

この章では，テキスト解析によってウェブページから解答群の作成を行なう手法について述べる．解答群の作成の流れを以下に示す．

1. 解答候補抽出
   (a) セグメント分割
   (b) セグメント検索
   (c) 解答候補抽出
2. 解答群の作成
   (a) 限定候補の抽出
   (b) 属性の付与
   (c) 解答群の作成
   (d) 解答群のスコア付け

これらの処理について，解答候補抽出の詳細を 5.1 節で述べ，解答群の作成の詳細を 5.2 節で述べる．

5.1　解答候補抽出

5.1.1　セグメント分割

ウェブページは一般に多くのトピックを含むため，ウェブページ全体から解答を探すのは効率が悪い．そこでウェブページをいくつかのセグメントに分割し，セグメントを単位として解答候補を抽出する．新聞記事を知識源とした先行研究では段落毎に文書を分割していた．しかしウェブページを知識源とした場合，段落を示すタグは存在するが，書き手によって段落の表現方法が異なるので段落の特定が難しい．そこでタグを用いてウェブページをセグメントに分割する．具体的には，h1，table，p のような，ウェブページにおいてある程度大きなまとまりを表わすタグやその境界となるタグを用いて分割を行なう．表 5.1 はセグメント分割に用いたタグの一覧である．これらのタグでウェブページを分割し，分割された個々の領域をセグメントとする．

表 5.1: セグメント分割に用いたタグ

blockquote, address, div, dl, hr, h1, h2, h3, h4, h5, h6, ol, ul, p, pre, noframes, noscript, table, dir, menu

5.1.2 セグメント検索

分割されたセグメントから解答が出現する可能性が高いセグメントを検索する．先行研究ではキーワードを全て含むセグメントを解答候補を抽出するセグメントとしていた．本研究ではこれに加え，あるキーワードがページのタイトルに含まれかつ残りのキーワードを全て含むセグメントも抽出する．ただし，プライマリキーワードは必ずセグメントに含まれなければならないとする．例えば,「アカデミー賞を受賞したのはだれですか」という質問の場合，プライマリキーワードは「受賞」，セカンダリキーワードは「アカデミー賞」となる．図 5.1 のように検索されたページのタイトルにセカンダリキーワードである「アカデミー賞」が含まれているならば，そのページ全体のトピックはアカデミー賞であると考えられる．そのウェブページのセグメントの中にプライマリキーワードである「受賞」が現れた場合は，それは「アカデミー賞」の受賞を指すと考えられる．また，キーワードのマッチングを行なう際には，Juman の代表表記を用いる．

図 5.1: 検索されるセグメントの例

5.1.3 解答候補抽出

セグメント検索で得られたセグメントを対象に解答候補を抽出する．ここでは抽出されたセグメントを入力とし，解答候補と解答候補のスコアを出力する．以下に解答候補抽出の処理の流れを示す．

1. 検索されたセグメントの解析
   形態素解析，構文解析をする．形態素解析には茶筌 [10]，構文解析には南瓜 [11] を用いる．

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．

条件 1：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いずれかの条件を満たす名詞であれば解答候補とする．表 5.2 に解答タイプごとに満たすべき条件を示す．

表 5.2: 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⟩		
数値		名詞-数 名詞-数+名詞-接尾-助数詞 名詞-数+記号-アルファベット	
その他		名詞全般	

条件 2：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 5.3 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 5.3 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わす.「解答候補のスコア」については後述する．

3. 解答候補のスコア
   解答候補は一般に複数得られるため，解答候補がどれだけ解答としてふさわしいか

表 5.3: 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

表わすスコアを付けて順位付けする．解答候補のスコアは以下の 3 つの要素によって決まる．

- 構文パターン
  解答候補を抽出するための構文パターンに関するスコア
- 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

解答候補のスコアの算出式を式 (5.1) に示す．

$$S_{ans} = w_{pat} \times S_{pat} + w_{mor} \times S_{mor} + w_{dis} \times S_{dis} \tag{5.1}$$

構文パターンに対するスコア $(w_{pat} \times S_{pat})$

表 5.3 に示した解答候補を抽出するための構文パターンに応じて決められるスコアである．スコアの詳細は表 5.3 に示す．また，w_{pat} はスコアに対する重みを表している．本研究では 0.4 とした．

品詞パターンに対するスコア $(w_{mor} \times S_{mor})$

解答候補を含む文を茶筌，南瓜を用いて形態素解析，固有表現タグ付けを行い，質問の解答タイプと一致する情報のタイプ (固有表現タグ，品詞，カタカナ) に応じた品詞情報のスコアを解答候補に与える．スコアの詳細を表 5.4 に示す．また，w_{mor} はスコアに対する重みを表している．本研究では 0.1 とした．

表 5.4: 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

距離情報に対するスコア ($w_{dist} \times S_{dist}$)

各キーワードと解答候補の近さに応じたスコアを計算する．距離スコアの算出式を式 (5.2) 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．

$$S_{dist} = \frac{1}{K} \sum_{j=1}^{K} \left(\frac{1}{dist(a_i, k_j) + 1} \times sent(a_i, k_j) \right) \tag{5.2}$$

- $K \cdots$総キーワード数
- $a_i \cdots$解答候補
- $k_j \cdots$キーワード
- $dist(a_i, k_j) \cdots a_i$と k_jの距離
- $$sent(a_i, k_j) = \begin{cases} 0.2 & (a_i \text{と } k_j \text{が 1 文中に存在しない場合}) \\ 1 & (a_i \text{と } k_j \text{が 1 文中に存在する場合}) \end{cases}$$

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，そのままでは分母が 0 になるため，分母に 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 (5.2) の $sent(a_i, k_j)$ によって実現されている．解答候補を含むセグメントを探す際，質問文中の全てのキーワードを含むセグメントだけでなく，セカンダリキーワードの一部はページのタイトルにあり，残りのキーワードの全てがセグメントにある場合でも，解答候補を取り出すセグメントとしている．このとき，タイトルに含まれるキーワードについては，式 (5.2) のスコアの計算に用いていない．つまり，式 (5.2) 中の K は同じセグメント内にあるキーワードの数を表わす．w_{dist} は距離情報に対するスコアの重みを表している．本研究では 0.5 とした．

5.2 解答群の作成

5.2.1 限定表現候補の抽出

解答候補を含むセグメントに含まれるキーワードに対し，その意味を限定する限定表現の候補を抽出する．限定表現を抽出するために 6 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，またはかぎ括弧で囲まれた表現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，セグメントを茶筌によっ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新聞記事コーパスにおける文書内の共起頻度に基づき，式 (5.3) の Dice 係数によって定義する．

$$D(x, y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5.3}$$

  - $|X| \cdots$名詞 x が出現する記事数
  - $|Y| \cdots$名詞 y が出現する記事数
  - $|X \cap Y| \cdots$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，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．本研究では閾値を 0.04 としている．

- セグメントの先頭 (s_{front})
  セグメントの先頭にはそのトピックを表わす場合が多い．そのため，セグメントの先頭が名詞ならば限定表現として抽出する．セグメントの先頭の名詞は，セグメントを茶筌によって形態素解析することで抽出する．

5.1.2 項で述べたように，キーワードにはセグメントに存在するものと，ページのタイトルに存在するものがある．セグメントに存在するキーワードの場合は上記 6 種類の限定表現を抽出し，タイトルに存在するキーワードの場合は,「近傍」と「セグメントの先頭」の 2 種類の限定表現を抽出する．両者の違いはプログラムの実装上の都合によるものである．本来はタイトルに含まれるキーワードに対しても，連体修飾，直前，直後，デ格のパターンを用いて限定表現を抽出するべきである．

5.2.2 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でなければ，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 5.2.1 項で抽出された (「解答」,「キーワード」,「限定表現」) という 3 つ組の集合から，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 9 つの種類がある．

- 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．

- 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:ー」，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．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」』とあった場合,「ディスカバリー」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 [13] を使用

した．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わす．

- 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:スポーツ」をたどり，このクラスを「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)
  5.2.1 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}) とセグメントの先頭 (s_{front}) は解答群をまとめる属性とはしない．近傍やセグメントの先頭に位置するという理由で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．したがって，ここでは以下の 4 つを限定表現の共通属性として用いる．

  - 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．

5.2.3 解答群の作成

5.2.1 項で述べた「解答候補 (a_i)」,「キーワード (k_j)」,「限定表現 (s_k)」の集合と 5.2.2 項で述べた「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．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．ただし，以下の場合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既に解答が 1 つに決まっているので，曖昧性がない．よって，曖昧性を表わす解答群としてふさわしくない．

- 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，質問の曖昧性を表わす解答群とみなすことはできない．よって，そのような場合は解答群は生成しない．

5.2.4 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．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，複数の解答群にスコア付けをする．解答群に対するスコアを式 (5.4) 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．

$$Score(AG(k_j, attr)) = w_1G + w_2A + w_3Z + w_4AZ + w_5D + w_6K \tag{5.4}$$

$$G = \frac{G_{type}}{|GG|}, A = \frac{A_{type}}{|AG|}, AZ = \frac{\sum_{a \in AG} az(a)}{|AG|}, D = \frac{\sum_{key,gen \in AG} dice(key, gen)}{|GG| \times |key|}, K = \frac{\sum_{a \in AG} k(a)}{|GG|} \tag{5.5}$$

- $G_{type} \cdots$ 解答群中の限定表現の異なり数
- $|GG| \cdots$ 解答群中の限定表現の数
- $A_{type} \cdots$ 解答群中の解答の異なり数
- $|AG| \cdots$ 解答群中の解答の数
- $\sum_{a \in AG} az(a) \cdots$ 解答群中の解答のスコアの総和
- $az(a) \cdots$ 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
- $\sum_{key,gen \in AG} dice(key, gen) \cdots$ 解答群中のキーワードと限定表現のダイス係数の総和
- $dice(key, gen) \cdots$ 解答群中のキーワードと限定表現のダイス係数
- $|key| \cdots$ キーワードの数
- $\sum_{a \in AG} k(a) \cdots$ 解答群中の限定表現の抽出パターンのスコアの総数
- $w_1, w_2, w_3, w_4, w_5, w_6 \cdots$ 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を意味する．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．このスコアが低い場合ほど同じ限定表現が異なる解答に対して出現していることを意味する．

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を意味する．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．このスコアが低い場合ほど同じ解答が異なる限定表現に対して出現していることを意味する．

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 5.5 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 5.5: 属性のスコア

属性		スコア
数量表現+接尾語 (N)		1
末尾 3 文字 ($E3$)		0.8
末尾 2 文字 ($E2$)		0.5
末尾 1 文字 ($E1$)		0.2
かぎ括弧 (K)		0.5
意味クラス (T)		0.6
上位の意味クラス (T')		0.6
限定表現の抽出パターン (S)	連体修飾 (s_{no})	0.3
	直前 (s_{prev})	0.3
	直後 (s_{succ})	0.3
	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．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．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ーンの平均スコア (K)

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ーンのスコアの総和である．表 5.6 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．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 5.6: 限定表現の抽出パターンのスコア

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})	0.3
セグメントの先頭 (s_{front})	0.3

なお，これらの重みやスコアの値は，スコア付けされた解答群の順位とスコアの詳細を，付録 A.1 の質問を用いた予備実験によって調べ，適切な解答群が上位に現れる値を設定した．

5.3 本手法と坂本の手法との比較

本節における手法は坂本らによって提案された手法を基盤としている．また，坂本らは解答を取り出す知識源として新聞記事を用いていたのに対し，本研究はウェブを知識源とする．このため，ウェブを知識源としたことにより，坂本らの手法をいくつかの点で改変した．ここではその改変の内容をまとめる．

文書分割

新聞記事を知識源とした坂本の研究では，文書を分割する際に新聞記事データに付与されている段落を示すタグに基づいて分割していた．ウェブページを知識源とした本研

究の場合，段落の表現方法が多様であることから，段落の特定が困難であるため，ウェブページにおいてある程度大きなまとまりやその境界となるタグに基づいて文書の分割を行なった．

セグメント検索

坂本の研究ではすべてのキーワードを含むセグメントのみを抽出していた．本研究では坂本のセグメント検索の条件に加え，あるキーワードがページのタイトルに含まれかつ残りのキーワードを全て含むセグメントも抽出した．セグメントがキーワード全てを含んでいない場合でも，タイトルにキーワードがある場合は，そのキーワードはページ全体のトピックを表わすと考えられるからである．

代表表記

本研究では，セグメント検索，解答候補の抽出，限定表現の抽出を行なう際のキーワードに Juman で用いられている代表表記を用いることで表記のゆれに対応し，柔軟にキーワードマッチングを行なった．

解答群のスコア付け

解答群のスコアを求める際，スコア付けのための重みをいくつか改変した．本研究ならびに坂本のシステムにおける，解答群のスコア付けの際に用いる属性のスコアを表 5.7，限定表現の抽出パターンのスコアを表 5.8 に示す．

属性のスコアの値は，最も適切な解答群が多かった数量表現と末尾 3 文字の属性のスコアを大きくするべきと考えた．しかし，これらに対するスコアは既に十分大きいため，意味クラスとかぎ括弧のスコアを減らすことで相対的にスコアを大きくした．また，ふさわしくない解答群が頻出した連体修飾，直前，直後のスコアは坂本のスコアよりもさらに低く設定した．

限定表現の抽出パターンのスコアの値は，ウェブページのように書き方が統率されていないテキストの場合，セグメントの先頭や近傍に位置する限定表現の候補は誤りである可能性が高いことを考慮し，坂本のスコアよりも低めに設定した．

表 5.7: 属性のスコアの比較

属性		坂本	本研究
数量表現+接尾語 (N)		1	1
末尾 3 文字 ($E3$)		0.8	0.8
末尾 2 文字 ($E2$)		0.5	0.5
末尾 1 文字 ($E1$)		0.2	0.2
かぎ括弧 (K)		0.7	0.5
意味クラス (T)		0.7	0.6
上位の意味クラス (T')		0.6	0.6
限定表現の抽出パターン (S)	連体修飾 (s_{no})	0.4	0.3
	直前 (s_{prev})	0.4	0.3
	直後 (s_{succ})	0.4	0.3
	デ格 (s_{de})	0.2	0.2

表 5.8: 限定表現の抽出パターンのスコアの比較

限定表現の抽出パターン	坂本	本研究
連体修飾 (s_{no})	1	1
直前 (s_{prev})	1	1
直後 (s_{succ})	1	1
デ格 (s_{de})	0.8	0.8
Dice 係数 (s_{dice})	0.5	0.3
セグメントの先頭 (s_{front})	0.5	0.3

第6章　評価実験

6.1　実験方法

4，5 章で述べた解答リストを獲得する手法を評価する実験を行なった．曖昧な質問を 30 個用意し，それに対して適切な解答リストが得られているかを評価した．評価に用いた質問の例と，その質問がどのように曖昧であるかの説明を表 6.1 に表わす．例えば「ノーベル賞を受賞した人はだれですか」という質問は,「部門」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，解答タイプ，キーワードタイプを表 6.2 に示すように人手で設定した．なお，評価で用いた全ての質問の一覧を付録 A.3 に，それらの解答タイプやキーワードの一覧は付録 A.4 に示す.

表 6.1: 質問の例

質問番号	質問文 何が曖昧か?
5	タイガースの監督はだれですか
	チーム (阪神，デトロイト)，日付
9	全英オープンで優勝したのはだれですか
	種目 (テニス，ゴルフ)，日付
17	ノーベル賞を受賞したのはだれですか
	部門，日付
26	オリンピックが開催されたのはどこですか
	日付
29	ワールドカップで優勝した国はどこですか
	日付

6.2　表の抽出の評価と考察

曖昧な質問に対する解答リストとして，4 章の手法を用いて抽出された表を評価した．抽出された表が質問に対する解答リストとして適切であるかを人手で判定した．表 6.3 に実験結果を示す．(A) はシステムが出力した表の総数である．(B) の適合率は，システム

表 6.2: 質問解析の結果

質問番号	プライマリキーワード	セカンダリキーワード	解答タイプ	キーワードタイプ
5	監督	タイガース	per	hyponym
9	優勝	全英オープン	per	agent
17	受賞	ノーベル賞	per	agent
26	開催	オリンピック	loc	agent
29	優勝	ワールドカップ	na	agent

が出力した表のうち適切な解答リストを含む割合である．(C) の再現率は以下のように求めた．今回の実験では，1 つの質問に対して検索結果の上位 100 件のウェブページを用いたため，合計 3000 個のウェブページが表の抽出処理の対象となる．これらのウェブページを人手で調べ，解答リストとしてふさわしい表を抽出した．再現率はこのようにして得られた表のうち，実際にシステムによって取り出された表の割合である．(D) は提案手法によって適切な解答リスト (表) が得られた質問の数を表わす.

表 6.3: 表の抽出の実験結果

(A) 抽出された表の数	24
(B) 適合率	46%
(C) 再現率	88%
(D) 解答リストが得られた質問の数	9

次に，この実験結果について，以下に例を挙げながら結果の分析と考察を述べる．

6.2.1 表の抽出の成功例

「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 の適切な解答リストを含む表が抽出された．ワールドカップには「回」,「開催年」等の曖昧性があるので適切な表であると判断した．この場合，曖昧なキーワードはワールドカップであり，その限定表現は「回」や「大会」である．本来なら表中の「回」が限定表現であることを特定し,「ワールドカップ」には開催された回数という観点で曖昧性があることを特定した上でユーザに提示するべきであるが，本研究では行っていない．また，提案手法によって抽出された正解の表を調べたところ，図 6.1 のような大会の開催回数や開催年に関する曖昧性が殆どであった．ワールドカップの例では，ラグビーやスキーといった競技に関しても曖昧性があるが，そのような観点でまとめた表は今回の実験では抽出できなかった．

30 個の質問に対して，正解リストを含む表が得られた質問の数は 9 であり，再現率が低いことから，提案手法は本来取り出すべき多くの表の抽出に失敗している．しかし，表

の抽出の精度は高く，誤った表を抽出した事例は3件であった．質問応答システムでは，正解となるすべての表を抽出する必要はなく，正しい表を1つだけ見つけてユーザに提示すれば十分であるので，再現率よりは精度が重視される．表6.3に示した結果は，上記のような観点からは望ましいと言える．

ページタイトル：サッカー　ワールドカップの歴史　第９回W杯

ワールドカップの歴史

回	大会	参加国	優勝	準優勝	3位
第1回	1930年ウルグアイ大会	13	ウルグアイ	アルゼンチン	ー
第2回	1934年イタリア大会	31	イタリア	チェコスロバキア	ドイツ
第3回	1938年フランス大会	36	イタリア	ハンガリー	ブラジル
第4回	1950年ブラジル大会	33	ウルグアイ	ブラジル	スウェーデン
第5回	1954年スイス大会	38	西ドイツ	ハンガリー	オーストリア
第6回	1958年スウェーデン大会	53	ブラジル	スウェーデン	フランス
第7回	1962年チリ大会	56	ブラジル	チェコスロバキア	チリ
第8回	1966年イングランド大会	71	イングランド	西ドイツ	ポルトガル
第9回	1970年メキシコ大会	71	ブラジル	イタリア	西ドイツ
第10回	1974年西ドイツ大会	98	西ドイツ	オランダ	ポーランド
第11回	1978年アルゼンチン大会	106	アルゼンチン	オランダ	ブラジル
第12回	1982年スペイン大会	109	イタリア	西ドイツ	ポーランド

図 6.1: 表の抽出の成功例

6.2.2　表の抽出の失敗例

適切な解答リストを含み，本来抽出するべき表の抽出に失敗した例を図6.2，6.3，6.4，6.5に示す．図6.2，6.3，6.4は「ノーベル賞を受賞したのは誰ですか」という質問に対する解答リストを含む表であり，図6.5は「全英オープンで優勝したのは誰ですか」という質問に対する解答リストを含む表である．表の抽出に失敗した要因は以下の通りである．

- キーワードと表の属性が一致していない．
  図6.2の例では，プライマリキーワードが「受賞」で表の属性が「氏名等」となっているため抽出に失敗している．

- 表に属性が存在しない．
  図6.3の例では，表内に解答の属性を表わすセル「受賞者」などが存在しないため抽出に失敗している．

- 属性が必ずしも1行目，1列目にない等の複雑な表に対して，キーワードと表の属性のマッチングに失敗する．
  図6.4の例では1行目に「日本人受賞者」というセルが存在するものの，提案手法では1行目，1列目の連結されているセルは無視し，次の行や列からマッチングを行なうという処理を行っているために，抽出に失敗している．

- 固有表現解析の失敗により，解答が並んだ行や列の認識に失敗する．
  図6.5の例では,「Tiger Woods」等の文字列の固有表現の解析に失敗している．

ページタイトル:京都大学-大学の紹介／概要　栄誉　ノーベル賞

受賞年	賞	氏名等	備考
2002年	化学賞	田中　耕一（東北大卒）	(株)島津製作所
	物理学賞	小柴　昌俊（東大卒）	東京大学名誉教授（受賞時）
2001年	化学賞	野依　良治（京大卒）	名古屋大学理学部 教授(受賞時)
2000年	化学賞	白川　英樹（東工大卒）	筑波大学名誉教授（受賞時）
1994年	文学賞	大江　健三郎（東大卒）	作家
1987年	医学・生理学賞	利根川　進（京大卒）	米国マサチューセッツ工科大学教授

図 6.2: 表の抽出の失敗例1

ページタイトル:ノーベル賞

1949年	物理学賞	湯川　秀樹
1965年	物理学賞	朝永　振一郎
1968年	文学賞	川端　康成
1974年	平和賞	佐藤　栄作
1973年	物理学賞	江崎　玲於奈
1981年	化学賞	福井　謙一

図 6.3: 表の抽出の失敗例2

6.2.3　誤って抽出した表の例

表の抽出は成功したが，誤った表を抽出してしまった例を図6.6，6.7に示す．図6.6は「ＮＨＫ杯で優勝したのは誰ですか」という質問に対してシステムが抽出した表であり，

ページタイトル：ノーベル賞

日本人受賞者						
	物理学賞	化学賞	生理学・医学賞	平和賞	文学賞	経済学賞
1949年	湯川　秀樹 博士					
1965年	朝永　振一郎 博士					
1968年					川端　康成 氏	
1973年	江崎　玲於奈 博士					
1974年				佐藤　栄作 首相		
1981年		福井　謙一 博士				

図 6.4: 表の抽出の失敗例 3

ページタイトル：Koogie Golf Town - Tiger Woods - The Open Championship Winners ★「全英オープン」優勝者とタイガーの成績（2007年～1995年）★ タイガー・ウッズ「全英オープン」優勝 ★Major★

「The Open Championship 全英オープン」優勝者				
No.	開催年	優勝者 Winners	開催地 Site	タイガー・ウッズ Tiger Woods
137	2008	☆	☆, ☆, ☆	☆位タイ
136	2007	Padraig Harrington	Carnoustie GC, Angus, UK	12位タイ
135	2006	Tiger Woods	Royal Liverpool Hoylake, England	優勝
134	2005	Tiger Woods	St. Andrews, Scotland	優勝
133	2004	Todd Hamilton	Royal Troon, Scotland	9位タイ
132	2003	Ben Curtis	Royal St. George's, England	4位タイ
131	2002	Ernie Els	Muirfield, Scotland	28位タイ

図 6.5: 表の抽出の失敗例 4

図 6.7 は「タイガースの監督は誰ですか」という質問に対してシステムが抽出した表である．失敗の要因は以下の通りである．

- 質問に対する解答を含まない表を抽出している．
  図 6.6 の例では，テレビアジア選手権に関する表であるのにも関わらず，1 行目に「優勝者」というセルがあり，表の前方の 3 セグメント以内に「ＮＨＫ杯」というセカンダリキーワードが存在する．このため，ＮＨＫ杯に関する表であるとみなしてしまい，誤って抽出している．

- 属性となるセルの誤検出．
  図 6.7 の例では，1 列目の「77-1 星野監督」というセルが，複合名詞であり，セルの末尾がプライマリキーワード「監督」となっているため，属性を表わすセルであるとみなしてしまった為に誤って表を抽出している

ページタイトル：棋戦情報|囲碁のポータルサイト|財団法人日本棋院

日本　趙治勲NHK杯(5)、結城聡九段(初)
韓国　李世ドル九段(2)、崔哲瀚九段(初)
中国　王檄九段(2)[前回優勝]、陳耀燁五段(初)、朴文尭五段(初)
テレビアジア選手権歴代優勝者

年	回	開催地	優勝者	代表	
1989	1回	日本	武宮　正樹	(日本)	☆
1990	2回	日本	武宮　正樹	(日本)	☆
1991	3回	韓国	武宮　正樹	(日本)	

図 6.6: 誤って抽出した表の例 1

6.3　テキスト解析による解答群の作成の評価と考察

曖昧な質問に対する解答リストとして，5 章で述べた手法を用いて作成した解答群を評価した．スコアの上位 10 件の解答群を人手で調べ，それが質問に対する解答リストとして適切であるかを判定した．ここで，解答群が適切であるということは，得られた解答群の限定表現と解答との対応が適切であり，限定表現が共通の属性で正しくまとめられていることを示す．また，ここでは解答候補を得てから解答群を生成するまでのアルゴリズムを評価する．このため，他のモジュールの誤りの影響を排除するため，解答候補抽出のステップで得られた解答のうち，質問に対して適切なもののみを用いて実験を行った．表 6.4 に実験結果を示す．(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質問の数とその割合である．(B) は最大のスコアを持つ解答群が解答リストとして適切であった質問の数とその割合である．(C) は適切な解答群を獲得できたが，それが最大のスコアを

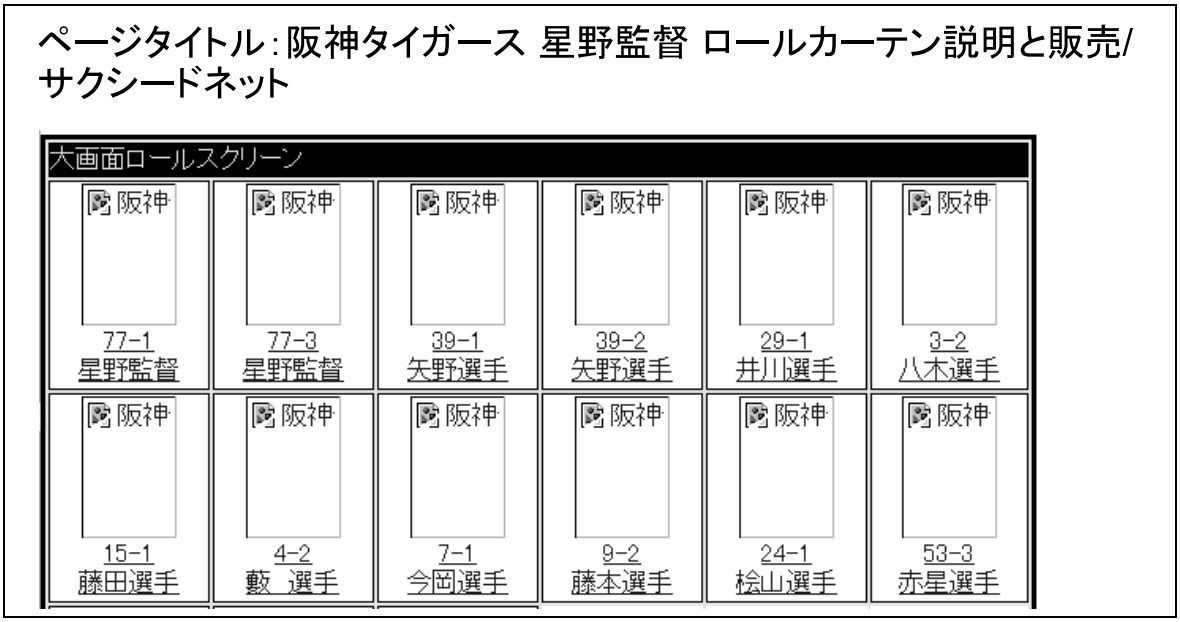

図 6.7: 誤って抽出した表の例 2

表 6.4: 解答群作成の実験結果

(A) 適切な解答群が存在しない	8(27%)
(B) 適切な解答群が得られ，かつその解答群が 1 位	13(43%)
(C) 適切な解答群が得られ，かつその解答群が 2 位〜10 位	9(30%)
(D) 適切な解答群が得られたときの平均順位	2.1 位

もつ解答群ではなかった質問の数とその割合である．(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位にあるかを表わす．

この実験結果について，以下に例を挙げながら結果の分析と考察を述べる．

6.3.1 解答群作成の成功例

「直木賞を受賞したのは誰ですか」という質問に対して図 6.8 の適切な解答群が作成された．直木賞には「回数」の曖昧性があるので適切な解答群であると判断した．図 6.8 の 1 行目「キーワード:直木賞」は「直木賞」というキーワードに関して曖昧性があるということを表わす.「属性:数量表現+接尾語:数+回」は，限定表現が「数量表現+接尾語」の共通属性，具体的には「数+回」という共通の属性を持つことを表わす.「score:0.6524」は解答群に対するスコア (式 (5.4)) を表わす．2 行目の「限定表現」はキーワード (図 6.8 の例では「直木賞」) の意味を限定する表現,「解答」は質問に対する解答を表わす．生成された解答群を調べたところ，6.2 節で抽出された表と同様に，大会の開催年や大会回数の曖昧性を表わすものが多かった．

キーワード:直木賞 属性:数量表現+接尾語:数+回 score:0.6524	
限定表現	解答
１３２回	角田光代
６９回	長部日出雄
１３３回	朱川湊人
１３７回	松井今朝子
５６回	五木寛之

図 6.8: 解答群作成の成功例

6.3.2 解答群作成の失敗例

30 個の質問中 8 個の質問は適切な解答群を得ることができなかった．その要因は以下のとおりである．

- 限定表現と解答との正しい対応が取れていない．
  図 6.9 は「水泳の世界選手権で優勝したのは誰ですか」という質問に対してシステムが作成した解答群である．この例では，北島康介は男子２００メートル平泳ぎでは優勝しているが，男子５０メートル平泳ぎでは優勝していない．同様に，ライアン・ロクテとマイケル・フェルプスも水泳の世界選手権の優勝者ではあるが，限定表現が表わす部門で優勝している訳ではない．セグメント内の近傍の単語や，セグメントの先頭から限定表現を抽出した場合にこのような失敗が起きる．

- 適切な限定表現がページ内に存在しない．
  例えば「社民党の党首は誰ですか」という質問がある．これは初代，二代目，三代目党首といった曖昧性や日本の社民党，ドイツの社民党という曖昧性があり，それぞれに党首が存在する．しかし，ウェブページ，もしくはセグメント自体に「三代目党首」.「ドイツ社民党」というような表記がなく，このような曖昧性を検出することができなかった．

キーワード:水泳 属性:数量表現+接尾語:数+メートル平泳ぎ score:0.4957	
限定表現	解答
男子２００メートル平泳ぎ	北島康介
男子５０メートル平泳ぎ	北島康介
男子２００メートル平泳ぎ	ライアン・ロクテ
男子５０メートル平泳ぎ	マイケル・フェルプス

図 6.9: 解答群作成の失敗例

6.4 組み合わせ手法の評価と考察

4 章と 5 章の手法を組み合わせた提案手法によって作成した解答リストを評価した．表 6.5 に実験結果を示す．表 6.5 における「表のみ」はウェブページから解答リストとなる表を抽出する手法 (4 章),「解答群のみ」は解答候補を抽出し，共通の属性を持つ限定表現によって解答群としてまとめた手法 (5 章),「併用」はまず表を抽出し，それに失敗した場合に解答群を作成するという方式で両者を併用する手法を表わす．一方，(A) は正解，不正解にかかわらず何かしらの解答リストが得られた質問の数，(B) は表ならば出力した表に正解を含む質問の数，解答群ならば 10 位以内に正解を含む質問の数を表わす，(C) は表ならば出力した表に正解を含む質問の数，解答群ならば 1 位に正解を含む質問の数を表わす．表 6.5 の結果から，(B)，(C) どちらの場合も 2 つの手法を組み合わせることで正解が得られる質問の数が増えていることがわかる．このことから，解答リストを得るために 2 つの手法を併用する提案手法は有効であるといえる．

表 6.5: 解答群作成の実験結果

	表のみ	解答群のみ	併用
(A) 出力全て	10	30	30
(B) 表+解答群 10 位以内に正解を含む	9	22	25
(C) 表+解答群 1 位に正解を含む	9	13	17

第7章　参照日付表現の処理

6章の考察で挙げたように，提案手法によって実際に出力される解答リストが示す質問の曖昧性として多いのは，年や日付に関連する曖昧性である．また，解答や限定表現としても日付表現が抽出される場合が多い．しかし,「今年」や「先月」のような相対的な日付表現が抽出された場合，これらをそのままユーザに提示しても，具体的な日時がわからない．そこで，相対的な日付表現を絶対的な日付に変換する必要がある．以下，相対的な日付表現のことを「参照日付表現」と呼ぶ.

先行研究のように新聞記事を知識源としている場合は，記事の日付を表わすタグに基づいて日付を抽出し，その日付を元に相対的な日付表現を変換すればよい．しかし，ウェブページには日付を記録するための統一された方法はなく，書き手によって日付の表現方法は多様である．また，日記やBlogのように，1つのウェブページ内においても違う日に書かれたテキストが混在する場合も多い．そこで，ウェブページ全体の更新日時やウェブページ内のある特定のテキストの更新日時を特定した上で，参照日付表現を変換する必要がある.

この章では参照日付表現を絶対的な日時に直すために，ウェブページにおけるセグメントが書かれた日付を特定する手法について検討する．ここで，セグメントとは，5.1節で述べた手法で検出されるウェブページの大きなまとまりを指す.

7.1　提案手法

セグメントが書かれた日付を特定するための手順を以下に示す.

1. 日付表現の抽出
2. セグメントとの対応づけ

これらの処理の詳細を以下に述べる.

日付表現の抽出

- ウェブページ全体の更新日時
  同じセグメント内に更新を示唆する表現と共に出現する日付表現を抽出する．このような日付表現はウェブページ全体が作成された日付を表わすことが多い.

- ウェブページの部分的な更新日時
  セグメントに単独で存在する日付表現を抽出する．このような日付表現は日記やBlog形式のウェブページにおいて，ウェブページの部分的なテキストの更新日時を表わす事が多い．

日付表現は以下のパターンで抽出した．また，年の表記がない日付表現も抽出した．

- 日本語の表記のパターン
  例：2008年2月7日，昭和58年11月13日
- 英語の表記のパターン
  例：2008,7,February，1983,13,Nov.
- 数字と記号による表記のパターン
  例：2008/02/07，2008/2/7，2008:11:13

次に，更新を示唆する表現を表7.1に示す．

表 7.1: 更新を示唆する表現

更新，更新日，更新日時，最終更新，最終更新日，最終更新日時，作成，作成日，作成日時，last update，update，updated，up，posted by，posted at

抽出される日付表現の例を示す．図7.1ではセグメント(A)内に更新を示唆する表現「最終更新」と共に出現する「2007年9月3日」と，セグメントに単独で存在する「2007年9月3日」,「2007年9月2日」,「2007年9月1日」の4つの日付表現が更新日時として抽出される．セグメント(B)内の「9月5日」という日付表現はセグメントに単独で出現しておらず，ウェブページ内のテキストの更新日時を表わしているわけではないので抽出しない．

セグメントとの対応付け

本研究では，ウェブページをセグメントに分割し，セグメントの中から解答候補や限定表現を抽出する．セグメント内にある参照日付表現を絶対的な日付に変換するためには，そのセグメントが書かれた日付を特定する必要がある．特に，日記やBlogのようにページ内に複数の更新日時が現われる場合，抽出された日付表現が解答候補を含むセグメントと正しく対応していなければならない．解答候補を含むセグメントと日付表現とを対応させる条件を以下に示す．

- 条件 1：セグメントと日付表現との距離が最も短い日付表現をセグメントの更新日時とする．
  一般に，日記や Blog ではセグメントに最も近い日付表現がそのセグメントの更新日時を表わす．ここでの距離とは，セグメントと日付表現の間に存在するセグメントの数と定義する．

- 条件 2：セグメントの前後の距離が等しい日付表現は，セグメントの前方に存在する日付表現を更新日時とする．
  距離が等しい日付表現が検出される場合の殆どは，人手で書かれた日記形式のページある．そのようなページの場合は，セグメントの前方に対応する日付表現が現れる．セグメントの後ろに対応する日付表現が現れる場合は殆どが Blog であり，その場合は条件 1 で検出される．

セグメントの対応付けの例を示す．図 7.1 のセグメント (B) の場合，距離が最も短くかつそのセグメントの前に現われる「2007 年 9 月 2 日」が更新日時となる．

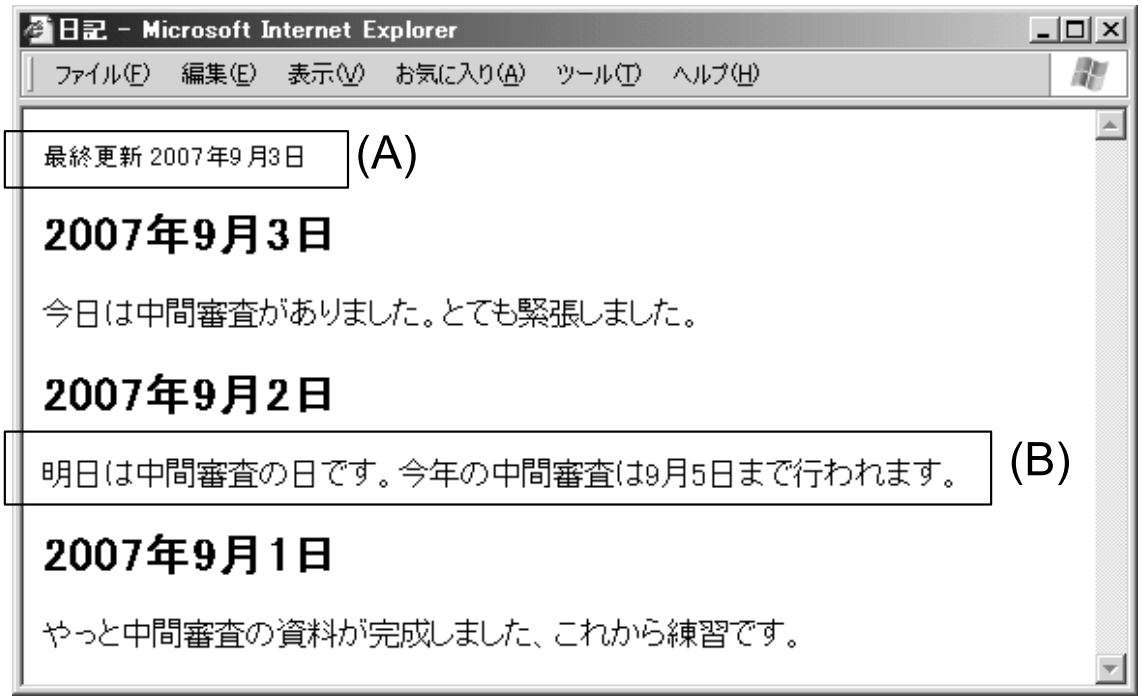

図 7.1: 日付表現の抽出の例

7.2 評価

7.1 節で述べたステップ 1 の日付表現を抽出する手法を評価する予備実験を行なった．この実験では，付録 A.1 に挙げた 25 個の質問を用意し，1 つの質問に対して TSUBAKI の

検索結果の上位 10 件，合計 250 個のウェブページを日付表現抽出の対象とした．抽出された日付表現が適切であるかを人手で判定した．適切な日付表現とは，抽出された日付表現が，ページ全体，もしくはページの部分的な更新日時を表わしている場合を指す．表 7.2 に実験結果を示す．(A) はシステムが出力した日付表現の総数である．(B) の適合率はシステムが出力した日付表現のうち適切な日付表現の割合である．(C) の再現率は本来抽出するべき日付表現のうち，システムが出力した適切な日付表現の割合である．

表 7.2: 日付表現抽出の実験結果

(A) 抽出された日付表現の数	30
(B) 適合率	67%
(C) 再現率	61%

本来抽出するべき日付表現の抽出に失敗した要因は,「二〇〇七年二月七日」や「2008/02/07 23:38」など 7.1 節に示した日付表現のパターンだけでは抽出できない日付表現が存在したことであった．

一方，日付表現の抽出は成功したが，誤った日付表現を抽出した要因は，人物のプロフィールなどに現れる生年月日やニュースサイトの記事の見出し等に含まれる日付表現であった．

また，今回の予備実験に用いたデータは，250 のウェブページに対して更新日時を表わす日付表現の数は 33 と少なく，提案手法の評価に適しているとはいえない．日付表現を多く含むページを対象に評価を行ない，またその結果を踏まえて日付表現を抽出する手法を改良する必要がある．

また，抽出した日付表現とセグメントの対応関係をとる処理や，参照日付表現を絶対的な日付に直す処理は実装していない．本論文では，参照日付表現を絶対的な日付表現に直す手法を検討しただけにとどまったが，提案システムに組み込むモジュールとして実装し，より適切な解答リストを提示できるようにすることが今後の課題である．

第8章　結論

本論文ではウェブを知識源とし，ユーザから曖昧な質問が入力されたとき，その質問の曖昧性を検出し，質問文中の曖昧なキーワードとその意味を限定する表現とともに複数の解答をリスト形式でユーザに提示するリスト型質問応答システムを提案した．曖昧な質問に対して知識源となるテキスト中の文を解析し，解答リストを動的に生成していた坂本らの手法を基盤とし，ウェブから適切な解答を得るために必要な改変に取り組んだ．これに加え，本論文ではウェブページにおける表に着目した．ユーザに提示する解答リストとなりうる表を発見する手法を提案し，従来のテキスト解析に基づく手法と併用した．

最後に本研究の今後の課題について述べる．

解答リストを含む表の抽出

- 多様な曖昧性の検出
  現在抽出されている表の多くは，賞の開催年や開催回数の曖昧性を反映したものであり，ユーザの多様な質問に対応できているとは言い難い．より多様な曖昧性を検出し，適切な表を抽出する手法を検討する必要がある．

- 表のランク付け
  現在，複数の表が抽出されたときはそのすべての表を出力している．複数の表をなんらかの基準でランク付けして，もっとも適切な表をひとつ選択する手法を検討する必要がある．

- 属性の検出
  現在属性の検出はプライマリキーワードとセルの文字列のマッチングによってのみ行っている．そこで人物を問う質問の場合は「氏名」，会社を問う質問の場合は「社名」を探すといったように，解答タイプに応じて適切な表の属性を検出する事で，表の抽出処理の再現率が向上すると考えられる．

テキスト解析による解答群の作成

- 解答候補抽出の改善
  文書を検索し，解答候補を抽出する際，解答を間違っ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．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．ウェブ文書は新聞記事とは異なり書き方が統率されていないため，現在の手法では抽出できない解答候補が多い．解答を抽出するパターンやキーワードの異表記に対応し，解答の抽出率を向上させる必要がある．

- 限定表現抽出の改善
  限定表現を抽出する際，限定表現を間違って取り出す場合や適切な限定表現を取り出せない場合がある．このことがふさわしくない解答群が生成される原因となっている．抽出された限定表現を取捨選択する手法や，ウェブページのタグや構造を考慮した限定表現を抽出する手法を検討する必要がある．

- 参照日付表現の対応
  現在，セグメントに対応する日付表現の抽出は行っているが，抽出した日付表現とセグメントの対応関係をとる処理や，参照日付表現を絶対的な日付に直す処理は実装していない．提案システムに組み込むモジュールとして実装し，より適切な解答リストを提示できるようにしたい．

謝辞

本研究を進めるにあたり，白井清昭 准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, 情報科学研究科, 情報処理学専攻, 修士論文,2006

[2] 坂本 篤史, 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, 情報科学研究科, 情報処理学専攻, 修士論文,2007

[3] Hideto Kazawa, Hideki Isozaki and Eisaku Maeda. NTT Question Answering System in TREC 2001. Text Retrieval Conference (TREC-2001), pp. 415-422, 2001.

[4] 石下円香, 森辰則. 優先順位型質問応答の解スコア分布に基づくリスト型質問応答. 情報処理学会自然言語処理研究会, Vol. 2005, No. 94, pp. 41–47, 2005.

[5] Jun'ichi Fukumoto,Tatsuhiro Niwa,Makoto Itogawa,Megumi Matsuda. Rits-QA:List answerdetection and Context task with ellipses handling. Working Notes of the forth NTCIR Workshop Meeting pp.310-314 June2004.

[6] 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

[7] 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

[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] 検索エンジン基盤 TSUBAKI http://tsubaki.ixnlp.nii.ac.jp/index.cgi

[10] 日本語形態素解析システム「茶筌」http://chasen.naist.jp/hiki/ChaSen/

[11] CaboCha/南瓜 http://www.chasen.org/ taku/software/cabocha/

[12] 日本語形態素解析システム JUMAN http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[13] 大野 晋, 浜西 正人. 「角川類語新辞典」角川書店,1981.

付 録A　実験に用いた質問一覧

表 A.1: 予備実験の質問一覧

質問番号	質問文 質問の曖昧な情報
1	府中市の市長はだれですか
	都道府県 (東京都，広島県)，日付
2	ＮＨＫ紅白歌合戦の司会はだれですか
	組 (白組，赤組)，日付
3	共産党の党首はだれですか
	国 (日本，中国，ソ連)，日付
4	自由民主党の党首はだれですか
	国 (日本，中国，ソ連)，日付
5	野球の首位打者はだれですか
	リーグ (セリーグ，パリーグ，大リーグ)，日付
6	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
7	野球連盟の理事長はだれですか
	団体 (大学野球，高校野球)，日付
8	野球の盗塁王はだれですか
	リーグ (セリーグ，パリーグ，大リーグ)，日付
9	サッカーの得点王はだれですか
	大会 (高校，Ｊリーグ，五輪，選手権)，日付
10	皐月賞で優勝したのはだれですか
	日付
11	新人王戦で優勝したのはだれですか
	種目 (囲碁，将棋)，日付
12	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
13	東カンファレンスで優勝したのはだれですか
	種目 (バスケット，アイスホッケー)，日付

質問番号	質問文 質問の曖昧な情報
14	ゴールデングローブ賞を受賞したのはだれですか
	部門 (投手部門，野手部門)，リーグ (セリーグ，パリーグ，大リーグ)，日付
15	日本有線大賞を受賞したのはだれですか
	部門，日付
16	天皇杯で優勝したのはどこですか
	種目 (サッカー，テニス，レスリング)，日付
17	全国高校ラグビーフットボール大会で優勝したのはどこですか
	日付
18	関東大学リーグで優勝したのはどこですか
	種目 (サッカー，アメリカンフットボール)，日付
19	東京箱根間往復大学駅伝競走で優勝したのはどこですか
	部門 (往路，復路，総合)，日付
20	カー，オブ，ザ，イヤーを受賞したのはどこですか
	地域 (日本，北米，欧州)，日付
21	新幹線が開業したのはいつですか
	地域 (東海道，山陽，東北)
22	ディズニーランドが開園したのはいつですか
	都市 (東京，アメリカ，香港)
23	ファイナルファンタジーが発売されたのはいつですか
	バージョン (7，8，9，X)
24	アジアカップで優勝した国はどこですか
	種目 (サッカー，卓球)，日付
25	万博が開催されたのはどこですか
	日付，回

表 A.2: 予備実験の質問設定一覧

質問番号	プライマリキーワード	セカンダリキーワード	解答タイプ	キーワードタイプ
1	市長	府中市	per	hyponym
2	司会	ＮＨＫ紅白歌合戦	per	hyponym
3	党首	共産党	per	hyponym
4	党首	自由民主党	per	hyponym
5	首位打者	野球	per	hyponym
6	歌手	入場行進曲，高校野球大会	per	hyponym
7	理事長	野球連盟	per	hyponym
8	盗塁王	野球	per	hyponym
9	得点王	サッカー	per	hyponym
10	優勝	皐月賞	per	agent
11	優勝	新人王戦	per	agent
12	優勝	全日本学生スキー選手権	per	agent
13	優勝	東カンファレンス	per	agent
14	受賞	ゴールデングローブ賞	per	agent
15	受賞	日本有線大賞	per	agent
16	優勝	天皇杯	org	agent
17	優勝	全国高校ラグビーフットボール大会	org	agent
18	優勝	関東大学リーグ	org	agent
19	優勝	東京箱根間往復大学駅伝競走	org	agent
20	受賞	カー・オブ・ザ・イヤー	org	agent
21	開業	新幹線	time	other
22	開園	ディズニーランド	time	other
23	発売	ファイナルファンタジー	time	other
24	優勝	アジアカップ	na	agent
25	開催	万博	loc	agent

表 A.3: 評価実験の質問一覧

質問番号	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	シドニー五輪の柔道の金メダリストはだれですか
	階級
3	日本の首相はだれですか
	日付，代
4	ジャイアンツの監督はだれですか
	チーム (読売，サンフランシスコ)，日付
5	タイガースの監督はだれですか
	チーム (阪神，デトロイト)，日付
6	野球の新人王はだれですか
	リーグ (セリーグ，パリーグ，大リーグ)，日付
7	社民党の党首はだれですか
	国 (日本，ドイツ)，日付，代
8	棋聖戦で優勝したのはだれですか
	種目 (囲碁，将棋)，日付
9	全英オープンで優勝したのはだれですか
	種目 (テニス，ゴルフ)，日付
10	全米オープンで優勝したのはだれですか
	種目 (テニス，ゴルフ)，日付
11	ＮＨＫ杯で優勝したのはだれですか
	種目 (フィギュアスケート，体操，将棋)，日付
12	日本グランプリで優勝したのはだれですか
	種目 (Ｆ１，オートバイ)，日付
13	水泳の世界選手権で優勝したのはだれですか
	種目，日付
14	グランドスラムを達成したのはだれですか
	種目 (テニス，ゴルフ，囲碁)，日付
15	アカデミー賞を受賞したのはだれですか
	部門，日付
16	レコード大賞を受賞したのはだれですか
	部門，日付
17	ノーベル賞を受賞したのはだれですか
	部門，日付

質問番号	質問文 質問の曖昧な情報
18	グラミー賞を受賞したのはだれですか
	部門，日付
19	芥川賞を受賞したのはだれですか
	日付
20	直木賞を受賞したのはだれですか
	日付
21	日本シリーズで優勝したのはどこですか
	種目 (野球，ゴルフ，将棋)，日付
22	六大学野球で優勝したのはどこですか
	地域 (東京，関西)，時期 (春季，秋季)
23	甲子園で優勝したのはどこですか
	時期 (春，夏)，日付
24	クラブユースサッカー選手権で優勝したのはどこですか
	カテゴリ (Ｕ－１８，Ｕ－１５)，日付
25	ワールドカップが開催されたのはどこですか
	日付
26	オリンピックが開催されたのはどこですか
	日付
27	サリン事件が起きたのはいつですか
	名前 (地下鉄，松本)
28	ドラゴンクエストが発売されたのはいつですか
	バージョン (１，２，３，４，５)
29	ワールドカップで優勝した国はどこですか
	日付
30	高尾山の標高はいくつですか
	都道府県 (東京，神奈川，石川)

表 A.4: 評価実験の質問設定一覧

質問番号	プライマリキーワード	セカンダリキーワード	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	per	hyponym
2	金メダリスト	シドニー五輪，柔道	per	hyponym
3	首相	日本	per	hyponym
4	監督	ジャイアンツ	per	hyponym
5	監督	タイガース	per	hyponym
6	新人王	野球	per	hyponym
7	党首	社民党	per	hyponym
8	優勝	棋聖戦	per	agent
9	優勝	全英オープン	per	agent
10	優勝	全米オープン	per	agent
11	優勝	ＮＨＫ杯	per	agent
12	優勝	日本グランプリ	per	agent
13	優勝	世界選手権，水泳	per	agent
14	達成	グランドスラム	per	agent
15	受賞	アカデミー賞	per	agent
16	受賞	レコード大賞	per	agent
17	受賞	ノーベル賞	per	agent
18	受賞	グラミー賞	per	agent
19	受賞	芥川賞	per	agent
20	受賞	直木賞	per	agent
21	優勝	日本シリーズ	org	agent
22	優勝	六大学野球	org	agent
23	優勝	甲子園	org	agent
24	優勝	クラブユースサッカー選手権	org	agent
25	開催	ワールドカップ	loc	agent
26	開催	オリンピック	loc	agent
27	起きた	サリン事件	time	other
28	発売	ドラゴンクエスト	time	other
29	優勝	ワールドカップ	na	agent
30	標高	高尾山	am	hyponym