(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4849596号
(P4849596)

(45)発行日　平成24年1月11日(2012.1.11)　(24)登録日　平成23年10月28日(2011.10.28)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 180A
G06F 17/30 320D
G06F 17/30 210A

請求項の数 10　(全 62 頁)

(21)出願番号　特願2005-354207(P2005-354207)
(22)出願日　平成17年12月8日(2005.12.8)
(65)公開番号　特開2007-157006(P2007-157006A)
(43)公開日　平成19年6月21日(2007.6.21)
審査請求日　平成20年11月14日(2008.11.14)

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100094662
弁理士　穂坂　和雄
(74)代理人　100096530
弁理士　今村　辰夫
(74)代理人　100119161
弁理士　重久　啓子
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　馬　青
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】質問応答装置、質問応答方法および質問応答プログラム

(57)【特許請求の範囲】
【請求項１】
第１のキーワード自体が複数のキーワードであり，かつ，第２のキーワード自体が複数のキーワードであり，第１のキーワード，第２キーワードが入力される構成であり，第１のキーワードと第２のキーワードから構成される自然言語で表現された質問データに対する解答とともに，第１のキーワードを増加して得る第３のキーワードと，第２のキーワードを増加して得る第４のキーワードとから構成される自然言語で表現された質問データに対する解答を出力する質問応答装置であって，
複数のキーワードが入力キーワードとして入力されるキーワード入力手段と，
前記入力キーワードに基づいて，前記入力キーワードの数より多いキーワードを抽出して出力キーワードとして出力するキーワード増加手段と，
前記キーワード増加手段により第１のキーワードを入力キーワードとして用いて増加して得た出力キーワードである前記第３のキーワードと，前記キーワード増加手段により第２のキーワードを入力キーワードとして用いて増加して得た出力キーワードである前記第４のキーワードとによって構成される質問に対する解答の候補である解答候補を，予め記憶された解答候補の検索対象である文書データ群から抽出する解答候補抽出手段と，
前記抽出された各解答候補が質問と対応付けられた表を解答表として出力する解答表出力手段とを備え，
前記キーワード増加手段は，
前記入力キーワードをキーワード抽出用の文書データが格納されたキーワード抽出用デ

ータベースで全文検索し，前記入力キーワードのうちの複数のキーワードの検索結果において前記複数のキーワードの前後に出現する文字列をパターンとして抽出するパターン抽出手段と，

前記パターン抽出手段で抽出したパターンを前記キーワード抽出用データベースで全文検索し，前記パターンに囲まれた表現を抽出し，前記抽出した表現を出力キーワードとして出力するキーワード抽出手段とを備える，

ことを特徴とする質問応答装置。

【請求項２】

請求項１に記載の質問応答装置において，

前記キーワード増加手段は，

前記入力された第１のキーワードに基づいて，第３のキーワードを出力キーワードとして出力し，前記入力された第２のキーワードに基づいて，第４のキーワードを出力キーワードとして出力し，

前記解答候補抽出手段は，予め用意された問題とその問題に対する解答の組の多数のセットを用いて，どういう問題のときにどういう解答になるかを学習し，その学習結果に基づいて，前記出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答の候補である解答候補を抽出する

ことを特徴とする質問応答装置。

【請求項３】

請求項１に記載の質問応答装置において，

前記キーワード増加手段は，

前記入力された第１のキーワードに基づいて，第３のキーワードを出力キーワードとして出力し，前記入力された第２のキーワードに基づいて，第４のキーワードを出力キーワードとして出力し，

前記解答候補抽出手段は，予め記憶手段中に格納された大量の文書データ群中から前記出力された第３のキーワードと第４のキーワードを含む文書データを取り出し，取り出された文書データの言語表現から，前記大量の文書データ群中に出現する頻度を用いて，前記出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出する

ことを特徴とする質問応答装置。

【請求項４】

請求項１に記載の質問応答装置において，

前記第２のキーワードに対応付けられた疑問代名詞が入力される疑問代名詞入力手段と，

前記疑問代名詞入力手段により入力された疑問代名詞に基づいて，前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え，

前記キーワード増加手段は，前記入力された第１のキーワードに基づいて，第３のキーワードを出力キーワードとして出力し，前記入力された第２のキーワードを出力キーワードとして出力し，

前記解答候補抽出手段は，前記解答候補の検索対象である文書データ群から，前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し，この検索処理で抽出された文書データから，前記解答タイプ推定手段によって推定された解答タイプに適合する言語表現を，前記第３のキーワードと第２のキーワードとによって構成される質問の解答候補として抽出する

ことを特徴とする質問応答装置。

【請求項５】

請求項１に記載の質問応答装置において，

予め定められた前記第２のキーワードに対応付けられた疑問代名詞に基づいて，前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解

答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え，
前記キーワード増加手段は，前記入力された第１のキーワードに基づいて，第３のキーワードを出力キーワードとして出力し，前記入力された第２のキーワードを出力キーワードとして出力し，
前記解答候補抽出手段は，前記解答候補の検索対象である文書データ群から，前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し，この検索処理で抽出された文書データから，前記解答タイプ推定手段によって推定された解答タイプに適合する言語表現を，前記第３のキーワードと第２のキーワードとによって構成される質問の解答候補として抽出する
ことを特徴とする質問応答装置。

【請求項６】
請求項１に記載の質問応答装置において，
前記キーワード増加手段は，
前記入力された第１のキーワードに基づいて，第３のキーワードを出力キーワードとして出力し，前記入力された第２のキーワードに基づいて，第４のキーワードを出力キーワードとして出力し，
前記解答候補抽出手段は，前記解答候補の検索対象である文書データ群から，前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードとを含む文書データを検索し，この検索処理で抽出された文書データから，予め定められた解答タイプに適合する言語表現を，前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出する
ことを特徴とする質問応答装置。

【請求項７】
請求項１に記載の質問応答装置において，
前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプであって，前記キーワード入力手段に入力された第２のキーワードに対応付けられた解答タイプが入力される解答タイプ入力手段を備え，
前記キーワード増加手段は，前記入力された第１のキーワードに基づいて，第３のキーワードを出力キーワードとして出力し，前記入力された第２のキーワードに基づいて，第４のキーワードを出力キーワードとして出力し，
前記第２のキーワードのうち前記出力された第４のキーワードに類似するものを，前記第４のキーワードのそれぞれについて，類似キーワードとして決定する類似キーワード決定手段を備え，
前記解答候補抽出手段は，前記解答候補の検索対象である文書データ群から，前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードを含む文書データを検索し，この検索処理で抽出された文書データから，前記出力された第４のキーワードが類似する類似キーワードに対応付けられて前記解答タイプ入力手段に入力された解答タイプに適合する言語表現を，前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出する
ことを特徴とする質問応答装置。

【請求項８】
請求項７に記載の質問応答装置において，
前記類似キーワード決定手段は，
予め記憶手段内に格納された大量の文書データ群中から，前記キーワード抽出手段によって出力された第４のキーワードと共起して出現する語である共起語を抽出するとともに，前記第４のキーワードのそれぞれについて，前記抽出された各共起語と共起して前記文書データ群中に出現する回数を要素とするベクトルである共起ベクトルを求め，
各第４のキーワードについての共起ベクトルと前記キーワード入力手段に入力された第２のキーワードと同一の第４のキーワードについての共起ベクトルとの類似の度合いを求

め，求められた類似の度合いに基づいて決まる，前記各第４のキーワードと類似する第２のキーワードと同一の第４のキーワードを，前記類似キーワードとする
ことを特徴とする質問応答装置。
【請求項９】
第１のキーワード自体が複数のキーワードであり，かつ，第２のキーワード自体が複数のキーワードであり，第１のキーワード，第２キーワードが入力される構成であり，第１のキーワードと第２のキーワードから構成される自然言語で表現された質問データに対する解答とともに，第１のキーワードを増加して得る第３のキーワードと，第２のキーワードを増加して得る第４のキーワードとから構成される自然言語で表現された質問データに対する解答を出力する質問応答方法であって，
複数のキーワードで構成される第１のキーワードと複数のキーワードで構成される第２のキーワードとを入力するステップと，
入力された前記第１のキーワードと第２のキーワードに基づいて，前記第１のキーワードと第２のキーワードを構成するそれぞれのキーワードの数を増加した第３のキーワードと第４のキーワードを抽出して出力するステップと，
前記第３のキーワードと第４のキーワードを含む増加した複数の出力キーワードによって構成される質問に対する解答の候補である解答候補を，予め記憶された解答候補の検索対象である文書データ群から抽出するステップと，
前記抽出された各解答候補と質問とが対応付けられた表を解答表として出力するステップとを有し，
前記入力キーワードの数より多いキーワードを抽出するステップは，
前記入力キーワードをキーワード抽出用の文書データが格納されたキーワード抽出用データベースで全文検索し，前記入力キーワードのうちの複数のキーワードの検索結果において前記複数のキーワードの前後に出現する文字列をパターンとして抽出するステップと，
前記抽出したパターンを前記キーワード抽出用データベースで全文検索し，前記パターンに囲まれた表現を抽出し，前記抽出した表現を出力キーワードとして出力するステップとを備える，
ことを特徴とする質問応答方法。
【請求項１０】
第１のキーワード自体が複数のキーワードであり，かつ，第２のキーワード自体が複数のキーワードであり，第１のキーワード，第２キーワードが入力される構成であり，第１のキーワードと第２のキーワードから構成される自然言語で表現された質問データに対する解答とともに，第１のキーワードを増加して得る第３のキーワードと，第２のキーワードを増加して得る第４のキーワードとから構成される自然言語で表現された質問データに対する解答を出力する質問応答装置が備えるコンピュータに実行させるための質問応答プログラムであって，
前記コンピュータに，
複数のキーワードで構成される第１のキーワードと複数のキーワードで構成される第２のキーワードとを入力する処理と，
入力された前記第１のキーワードと第２のキーワードに基づいて，前記第１のキーワードと第２のキーワードを構成するそれぞれのキーワードの数を増加した第３のキーワードと第４のキーワードを抽出して出力する処理と，
前記第３のキーワードと第４のキーワードを含む増加した複数の出力キーワードによって構成される質問に対する解答の候補である解答候補を，予め記憶された解答候補の検索対象である文書データ群から抽出する処理と，
前記抽出された各解答候補と質問とが対応付けられた表を解答表として出力する処理とを実行させると共に，
前記入力キーワードの数より多いキーワードを抽出する処理は，
前記入力キーワードをキーワード抽出用の文書データが格納されたキーワード抽出用デ

ータベースで全文検索し，前記入力キーワードのうちの複数のキーワードの検索結果において前記複数のキーワードの前後に出現する文字列をパターンとして抽出する処理と，
前記抽出したパターンを前記キーワード抽出用データベースで全文検索し，前記パターンに囲まれた表現を抽出し，前記抽出した表現を出力キーワードとして出力する処理とを実行させる
ことを特徴とする質問応答プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、コンピュータによる自然言語処理システムとして、自然言語で表現された質問に対する解答を出力する質問応答技術に関し、特に、入力されたキーワードをキーワード抽出技術によって増加させ、増加したキーワードによって構成される複数の質問に対する解を自動的に求めて出力する質問応答装置、質問応答方法および質問応答プログラムに関する。

【０００２】

質問応答装置とは、自然言語による質問を入力すると、その解答そのものを出力する装置である。例えば、「パーキンソン病の兆候は脳のどの部分にある細胞の死が関係していますか。」という質問を入力すると、Ｗｅｂ、新聞記事、事典などのデータを含む大量の電子化テキストから「パーキンソン病は、中脳の黒質にあるメラニン細胞が変性し、黒質細胞内で作られる神経伝達物質のドーパミンがなくなり発病する、とされている。」といった文を探し出し、「黒質」と的確に解答を出力する。

【０００３】

質問応答装置は、論理式やデータベースからではなく、自然言語で記述された普通の文（テキストデータ）から解答を取り出すことができるため、大量の既存の文書データを利用することができる。また、質問応答装置は、キーワードで検索された記事から使用者自らが解答を探す必要がある情報検索システムなどと異なり、解答自体を出力する。そのため、使用者は、より早く解答の情報を得ることができる。このように質問応答装置は有用であるため、より使いやすい実用的な質問応答装置の実現が期待されている。

【背景技術】

【０００４】

一般的な質問応答装置（または質問応答システム）は、おおまかに、解答表現推定処理、文書検索処理、解答抽出処理という３つの処理手段で構成されている。

【０００５】

解答表現推定処理は、入力した質問中の疑問代名詞の表現などに基づいて解答表現を推定する処理である。解答表現とは、所望される解答の言語表現の類型であって、解答となる言語表現の意味に基づいた類型（解答タイプ）、解答となる言語表現の表記に基づいた類型（解答表現タイプ）などがある。質問応答装置は、どのような質問の言語表現がどのような解答表現を要求しているかという対応関係を参照して、入力した質問の解答の解答タイプを推定する。質問応答装置は、例えば、入力した質問が「日本の面積はどのくらいですか」である場合には、所定の対応関係を参照して、質問中の「どのくらい」という表現から解答タイプは「数値表現」であると推定する。また、質問が「日本の首相はだれですか」という場合には、質問中の「だれ」という表現から、解答タイプは「固有名詞（人名）」であると推定する。

【０００６】

文書検索処理は、質問からキーワードを取り出し、このキーワードを用いて解答を検索する対象となっている文書データ群を検索し、解答が記述されていると考えられる文書データを抽出する処理である。質問応答装置は、例えば、入力された質問が「日本の首都はどこですか」である場合に、質問から「日本」および「首都」をキーワードとして抽出し、検索対象の文書データ群から、キーワード「日本」および「首都」を含む文書データを検索する。

【０００７】
解答抽出処理は、文書検索処理で抽出されたキーワードを含む文書データから、推定した解答タイプに適合する言語表現を抽出し、解答として出力する処理である。質問応答装置は、例えば、文書検索処理において検索されたキーワード「日本」および「首都」を含む文書データから、解答表現推定処理において推定した解答タイプ「固有名詞（地名）」に適合する言語表現「東京」を抽出して解答とする。
【０００８】
前記のような処理を行うことにより、質問応答装置は、質問「日本の首都はどこですか」に対して解答「東京」を出力する。
【０００９】
なお、質問応答装置（または質問応答システム）に関する具体的な従来技術として、例えば、下記の非特許文献１に、複数の記事を使って解答の推定を行う質問応答システムにおいて、複数の記事から得られた解答の候補の得点を少しずつ減らしながら加算し、合計点が最も高い候補を解答として出力する技術について記載されている。
【非特許文献１】村田真樹，井佐原均，質問応答システムにおける逓減加点法に基づく複数記事情報の利用，情報処理学会自然言語処理研究会 2004-NL-160，2004年．九州大学．
【発明の開示】
【発明が解決しようとする課題】
【００１０】
従来の質問応答装置では、検索された文書データから解答となりうる言語表現を解答候補として抽出し、抽出した解答候補それぞれの解答タイプを判定する。そして、質問から推定した解答タイプと同じか類似する解答タイプと判定した解答候補の評価を高くし、原則的には、解答タイプが同じ解答候補であって所定の評価を得たものを解答として出力する。
【００１１】
しかし、従来の質問応答装置は、質問の入力によって問い合わせられた質問に対する解答のみを出力するシステムであって、問い合わせられた質問以外の質問に対する解答を出力することはできなかった。
【００１２】
本発明は、上記従来技術の問題点を解決し、問い合わせられた質問に対する解答および問い合わせられた質問以外の質問に対する解答を出力する質問応答装置、質問応答方法および質問応答プログラムの提供を目的とする。
【課題を解決するための手段】
【００１３】
上記課題を解決するため、本発明は、自然言語で表現された質問データに対する解答を出力する質問応答装置であって、複数のキーワードが入力キーワードとして入力されるキーワード入力手段と、前記入力キーワードに基づいて、前記入力キーワードの数より多いキーワードを抽出して出力キーワードとして出力するキーワード増加手段と、前記出力キーワードによって構成される質問に対する解答の候補である解答候補を、予め記憶された解答候補の検索対象である文書データ群から抽出する解答候補抽出手段と、前記抽出された各解答候補が質問と対応付けられた表を解答表として出力する解答表出力手段とを備えることを特徴とする。
【００１４】
また、本発明は、前記の質問応答装置において、前記キーワード増加手段は、前記入力キーワードをキーワード抽出用の文書データが格納されたキーワード抽出用データベースで全文検索し、前記入力キーワードの周辺に出現したパターンを抽出するパターン抽出手段と、前記パターン抽出手段で抽出したパターンを前記キーワード抽出用データベースで全文検索し、前記パターンによって抽出される表現を抽出し、前記抽出した表現を出力キーワードとして出力するキーワード抽出手段とを備えることを特徴とする。
【００１５】

また、本発明は、前記の質問応答装置において、前記キーワード増加手段は、前記入力キーワードと同じ分野の単語を、単語と単語の分野との対応情報が格納されたデータベースから抽出し、出力キーワードとして出力することを特徴とする。

【００１６】

また、本発明は、前記の質問応答装置において、前記キーワード増加手段は、予めデータベース中に記憶された、意味的類似による単語の分類情報であるシソーラスデータに基づいて、前記入力された入力キーワードと、前記シソーラスデータ中の単語との類似度を算出する類似度算出手段と、前記算出された類似度の大きさに基づいてキーワードを抽出し、出力キーワードとして出力するキーワード抽出手段とを備えることを特徴とする。

【００１７】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、予め用意された問題とその問題に対する解答の組の多数のセットを用いて、どういう問題のときにどういう解答になるかを学習し、その学習結果に基づいて、前記出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答の候補である解答候補を抽出することを特徴とする。

【００１８】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、予め記憶手段中に格納された大量の文書データ群中から前記出力された第３のキーワードと第４のキーワードを含む文書データを取り出し、取り出された文書データの言語表現から、前記大量の文書データ群中に出現する頻度を用いて、前記出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出することを特徴とする。

【００１９】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記第２のキーワードに対応付けられた疑問代名詞が入力される疑問代名詞入力手段と、前記疑問代名詞入力手段に入力された疑問代名詞に基づいて、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、前記解答タイプ推定手段によって推定された解答タイプに適合する言語表現を、前記第３のキーワードと第２のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２０】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、予め定められた前記第２のキーワードに対応付けられた疑問代名詞に基づいて、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワー

ドとして出力し、前記入力された第２のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、前記解答タイプ推定手段によって推定された解答タイプに適合する言語表現を、前記第３のキーワードと第２のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２１】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、疑問代名詞が入力される疑問代名詞入力手段と、前記疑問代名詞入力手段に入力された疑問代名詞に基づいて、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、前記解答タイプ推定手段によって推定された解答タイプに適合する言語表現を、前記第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２２】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、予め定められた疑問代名詞に基づいて、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、前記解答タイプ推定手段によって推定された解答タイプに適合する言語表現を、前記第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２３】

また、本発明は、前記の質問応答装置において、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプが入力される解答タイプ入力手段を備え、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、前記解答タイプ入力手段に入力された解答タイプに適合する言語表現を、前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２４】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キー

ワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、予め定められた解答タイプに適合する言語表現を、前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２５】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプであって、前記キーワード入力手段に入力された第２のキーワードに対応付けられた解答タイプが入力される解答タイプ入力手段を備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、前記解答タイプ入力手段に入力された解答タイプに適合する言語表現を、前記出力された第３のキーワードと第２のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２６】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードを出力キーワードとして出力し、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第２のキーワードとを含む文書データを検索し、この検索処理で抽出された文書データから、予め定められた、前記第２のキーワードに対応付けられた解答タイプに適合する言語表現を、前記出力された第３のキーワードと第２のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２７】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプであって、前記キーワード入力手段に入力された第２のキーワードに対応付けられた解答タイプが入力される解答タイプ入力手段を備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記第２のキーワードのうち前記出力された第４のキーワードに類似するものを、前記第４のキーワードのそれぞれについて、類似キーワードとして決定する類似キーワード決定手段を備え、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、前記出力された第４のキーワードが類似する類似キーワードに対応付けられて前記解答タイプ入力手段に入力された解答タイプに適合する言語表現を、前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２８】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入

力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記第２のキーワードのうち前記出力された第４のキーワードに類似するものを、前記第４のキーワードのそれぞれについて、類似キーワードとして決定する類似キーワード決定手段を備え、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、前記出力された第４のキーワードが類似する類似キーワードに予め対応付けられた解答タイプに適合する言語表現を、前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００２９】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、前記第２のキーワードに対応付けられた疑問代名詞が入力される疑問代名詞入力手段と、前記疑問代名詞入力手段に入力された疑問代名詞に基づいて、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記第２のキーワードのうち前記出力された第４のキーワードに類似するものを、前記第４のキーワードのそれぞれについて、類似キーワードとして決定する類似キーワード決定手段を備え、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、前記出力された第４のキーワードが類似する類似キーワードに対応付けられて前記疑問代名詞入力手段に入力された疑問代名詞に基づいて解答タイプ推定手段が推定した解答タイプに適合する言語表現を、前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００３０】

また、本発明は、前記の質問応答装置において、前記キーワード入力手段には、前記入力キーワードとして第１のキーワードと第２のキーワードとが入力され、予め定められた、前記第２のキーワードに対応付けられた疑問代名詞に基づいて、前記キーワード増加手段によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定手段とを備え、前記キーワード増加手段は、前記入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、前記入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力し、前記第２のキーワードのうち前記出力された第４のキーワードに類似するものを、前記第４のキーワードのそれぞれについて、類似キーワードとして決定する類似キーワード決定手段を備え、前記解答候補抽出手段は、前記解答候補の検索対象である文書データ群から、前記キーワード増加手段によって出力された第３のキーワードと第４のキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、前記出力された第４のキーワードが類似する類似キーワードに対応付けられた疑問代名詞に基づいて解答タイプ推定手段が推定した解答タイプに適合する言語表現を、前記出力された第３のキーワードと第４のキーワードとによって構成される質問の解答候補として抽出することを特徴とする。

【００３１】

また、本発明は、前記の質問応答装置において、前記類似キーワード決定手段は、予め記憶手段内に格納された大量の文書データ群中から、前記キーワード抽出手段によって出力された第４のキーワードと共起して出現する語である共起語を抽出するとともに、前記

第４のキーワードのそれぞれについて、前記抽出された各共起語と共起して前記文書データ群中に出現する回数を要素とするベクトルである共起ベクトルを求め、各第４のキーワードについての共起ベクトルと前記キーワード入力手段に入力された第２のキーワードと同一の第４のキーワードについての共起ベクトルとの類似の度合いを求め、求められた類似の度合いに基づいて決まる、前記各第４のキーワードと類似する第２のキーワードと同一の第４のキーワードを、前記類似キーワードとすることを特徴とする。

【００３２】

また、本発明は、前記の質問応答装置において、前記類似キーワード決定手段は、予めデータベース中に記憶された、意味的類似による単語の分類情報であるシソーラスデータに基づいて、前記キーワード増加手段によって出力された第４のキーワード毎に、前記第４のキーワードと同一の単語と、前記キーワード入力手段に入力された第２のキーワードと同一の単語との類似度を算出する類似度算出手段と、前記算出された類似度の大きさに基づいて決まる、前記第４のキーワードと類似する第２のキーワードを、前記類似キーワードとすることを特徴とする。

【００３３】

また、本発明は、自然言語で表現された質問データに対する解答を出力する質問応答方法であって、複数のキーワードを入力キーワードとして入力するステップと、前記入力キーワードに基づいて、前記入力キーワードの数より多いキーワードを抽出して出力キーワードとして出力するステップと、前記出力キーワードによって構成される質問に対する解答の候補である解答候補を、予め記憶された解答候補の検索対象である文書データ群から抽出するステップと、前記抽出された各解答候補が質問と対応付けられた表を解答表として出力するステップとを有することを特徴とする。

【００３４】

また、本発明は、自然言語で表現された質問データに対する解答を出力する質問応答装置が備えるコンピュータに実行させるためのプログラムであって、前記コンピュータに、複数のキーワードを入力キーワードとして入力する処理と、前記入力キーワードに基づいて、前記入力キーワードの数より多いキーワードを抽出して出力キーワードとして出力する処理と、前記出力キーワードによって構成される質問に対する解答の候補である解答候補を、予め記憶された解答候補の検索対象である文書データ群から抽出する処理と、前記抽出された各解答候補が質問と対応付けられた表を解答表として出力する処理とを実行させるための質問応答プログラムである。

【発明の効果】

【００３５】

本発明の質問応答装置によれば、問い合わせられた質問に対する解答だけでなく、問い合わせられた質問以外の質問に対する解答を、各質問に対応付けた形式で出力することが可能となる。すなわち、本発明の質問応答装置によれば、ユーザは、解答を知りたいジャンルのキーワードを少数入力するだけで、入力されたキーワードに基づいて増加したキーワードによって構成される多数の質問に対する解答を自動的に得ることができる。

【００３６】

例えば、本発明の質問応答装置によれば、ユーザが第１のキーワードと第２のキーワードとを入力すると、第１のキーワードに基づいて、第１のキーワードの数より多い第３のキーワードが抽出されるとともに、第２のキーワードに基づいて、第２のキーワードの数より多い第４のキーワードが抽出され、抽出された第３のキーワードと第４のキーワードに基づいて構成される質問に対する解答を機械学習の手法を用いて自動的に出力することが可能となる。

【００３７】

また、例えば、本発明の質問応答装置によれば、ユーザが第１のキーワードと第２のキーワードと、第２のキーワードに対応付けられた疑問代名詞とを入力すると、第１のキーワードの数より多い第３のキーワードが抽出されるとともに、上記入力された疑問代名詞に基づいて解答タイプが推定され、第３のキーワードと第２のキーワードと疑問代名詞に

基づいて構成される質問に対する解答を、上記推定された解答タイプを用いて自動的に出力することが可能となる。

【００３８】

また、例えば、本発明の質問応答装置によれば、ユーザが第１のキーワードと第２のキーワードと、解答タイプとを入力すると、第１のキーワードに基づいて、第１のキーワードの数より多い第３のキーワードが抽出されるとともに、第２のキーワードに基づいて、第２のキーワードの数より多い第４のキーワードが抽出され、抽出された第３のキーワードと第４のキーワードに基づいて構成される質問に対する解答を、上記入力された解答タイプを用いて自動的に出力することが可能となる。

【００３９】

また、例えば、本発明の質問応答装置によれば、ユーザが第１のキーワードと第２のキーワードと、第２のキーワードに対応付けられた解答タイプとを入力すると、第１のキーワードに基づいて、第１のキーワードの数より多い第３のキーワードが抽出されるとともに、第２のキーワードに基づいて、第２のキーワードの数より多い第４のキーワードが抽出され、さらに、抽出された第４のキーワードに類似する第２のキーワード（と同一の第４のキーワード）が類似キーワードとして決定される。そして、抽出された第３のキーワードと第４のキーワードに基づいて構成される質問に対する解答を、上記決定された類似キーワードに対応付けられた解答タイプを用いて自動的に出力することが可能となる。

【発明を実施するための最良の形態】

【００４０】

まず、本発明の実施の形態の説明の前に、上記非特許文献１に記載された技術について説明する。非特許文献１では、質問応答システムにおける逓減加点法に基づく複数記事情報の利用について記載されている。以下に非特許文献１の記載内容について説明する。

【００４１】

質問応答システムは、与えられた質問に対してその答えを出力するシステムのことで、例えば、「日本の首都はどこですか」という質問文が与えられると、「東京は日本の首都で、その国の最も大きく重要な都市であり、東京は日本の４７都道府県のうちの一つである。」という文をウェブや新聞記事などの電子テキストから探し出し、「東京」と答える。質問応答システムは、情報検索の代りとして重要になるだろうし、また将来の人工知能システムの基本要素にもなるであろう重要なものである。

【００４２】

非特許文献１では、質問応答システムの精度向上のために、複数の記事から得た解の候補の得点を減らしながら加点する新しい方法を提案している。この方法を逓減加点法と呼ぶ。

【００４３】

質問の答えが複数の記事で見つかることは多く、そのような場合は、複数の記事を使って答えを推定した方が一つの記事を使って推定するよりも良い答えを得ることができると思われるので、複数の記事から得た解の候補の得点を加算することで、複数の記事の情報を利用する手法が考えられる。しかし、ただ単純に得点を加算するだけではシステムの性能を下げる場合がある。

【００４４】

そこで、非特許文献１では、この単純に加算する際に生じる問題に対処するために、得点の加算の際に得点を減らしながら加算する手法を用いる。より具体的に言うと、非特許文献１の方法では、ｉ番目の解の候補の得点には$k^{(i-1)}$の重みをかけておいて、その後で得点を加算する。最終的な答えは合計得点により判断する。例えば、「東京」が三つの記事から解の候補として抽出され、それらの得点が２６、２１、２０であり、ｋが０．３であったとする。この場合、「東京」の合計得点は、３４．１となる（＝２６＋２１×０．３＋２０×0.3^2）。このような方法でそれぞれの候補の得点を計算し、最も高い合計得点を持つ候補を解とする。

【００４５】

次に、非特許文献１における複数記事の利用における逓減加点法の利用について詳細に説明する。「日本の首都はどこですか」という質問文が与えられたとする。このとき、得るべき答えは「東京」である。一般的な質問応答システムは、図２１のように、解の候補と得点をリストとして出力でき、また、解の候補を取り出した記事を指し示す記事番号も出力することができる。なお、図中に示す順位は、得点の大きさの順位を示す。

【００４６】

図２１に示すリストの例だと、最も得点の大きい候補は「京都」であり、誤った解を出力することになる。解の候補の得点を単純に加算する方法は、すでに提案されている。図２１に示すリストを用いると、解の候補の得点を単純に加算する方法によれば、図２２に示す結果を得る。

【００４７】

図２２では、「東京」の得点が一番順位が高く、システムは、正しく「東京」を解として出力することができる。この、解の候補の得点を単純に加算する方法は、複数の記事の情報を利用することで正しい解を得ることができた。しかし、この方法には、高頻度の解の候補を取り出しやすいという問題がある。これは、特に性能が高いシステムで深刻な問題である。もともと性能が高いシステムでは、システムの出力した元の得点の方が単純に加算した得点よりも信頼できる場合が多く、単純に加算する方法は、しばしばシステムの性能を劣化させることになる。

【００４８】

この問題に対処するために、非特許文献１の技術は、得点を減らしながら加算する新しい方法を提案している。解の候補の得点を単純に加算する代りに、得点を減らす重みをつけて得点を加算するのである。この方法は、高頻度語を取り出し易いという悪い効果を減じ、なおかつシステムの性能を向上させる効果を持つ。

【００４９】

この、非特許文献１で提案する方法の有効性を示す例をあげる。「日本の首都は西暦１０００年の時はどこでしたか。」と質問が与えられ、システムは図２３に示す結果を出力したとする。図２３に示すように、「京都」の得点が一番高い。ここで、上記質問に対する正解は「京都」であり、解の候補の得点を単純に加算しなければ、このシステムは正解を出力している。しかし、単純に加算する方法を用いると、その結果は図２４に示す表のようになり、間違った解の「東京」をシステムの解としてしまう。

【００５０】

ここで、得点を減らしながら加算する非特許文献１の新しい方法を利用してみる。ここでは、細かいシステムの仕様として、ｉ番目の候補の得点に$0.3^{(i-1)}$を乗じることとする。その場合、「東京」の得点は２．８であり（＝２．１＋１．８×０．３＋１．５×0.3^2＋１．４×0.3^3）、システムの出力結果は、図２５に示す表のようになり、「京都」の得点が一番高いので、正解の「京都」を解として正しく出力することができる。すなわち、非特許文献１で提案する方法は、最初の例（「日本の首都はどこですか」という質問文が与えられた場合）でも正しい解を得ることができる。最初の例に適用すると、「東京」の得点は４．３となり（＝３．２＋２．８×０．３＋２．５×0.3^2＋２．４×0.3^3）、出力結果は図２６に示す表のようになり、「東京」が最も高い得点となり、解として正しく出力される。

【００５１】

得点を減らしながら加算する非特許文献１に記載された方法は、高頻度の解の候補を取り出しやすい欠点を減じながら、なおかつ複数記事の情報を利用し精度向上を実現できるものである。

【００５２】

非特許文献１に記載された質問応答システムは、以下の三つの基本要素からなる。

１．解表現の推定

質問応答システムは、疑問代名詞の表現などに基づいて解表現（解がどのような言語表現か）を推定する。例えば、入力の質問文が「日本の面積はどのくらいですか」だとする

と、「どのくらい」という表現から、解表現は数値表現であろうと推測する。
２．文書検索
質問応答システムは、質問文からキーワードを取り出し、これらのキーワードを用いて文書を検索する。この検索により、解が書いてありそうな文書群を集めることになる。例えば、入力の質問文が、「日本の面積はどのくらいですか」だとすると、「日本」、「面積」がキーワードとして抽出され、これらを含む文書を検索することになる。
３．解の抽出
質問応答システムは、解が書いてありそうな文書群から、推定した解表現に適合する言語表現を抽出し、それを解として出力する。例えば、入力の質問文が、「日本の面積はどのくらいですか」だとすると、文書検索で検索した「日本」、「面積」を含む文書群から、解表現として推定した数値表現にあたる言語表現を解として抽出する。
【００５３】
以下に、非特許文献１で提案する技術について、詳細に説明する。
【００５４】
（解表現の推定）
人手で作成したヒューリスティックルールを使って解表現を推定する。１６個のルールを作成する。そのいくつかを以下に示す。
・質問文に「誰」という表現がある場合、解表現は人名である。
・質問文に「いつ」という表現がある場合、解表現は時間表現である。
・質問文に「どのくらいの」という表現がある場合、解表現は数値表現である。
【００５５】
（文書検索）
文書検索のためのキーワードは、公知のキーワード抽出ツールであるＣｈａＳｅｎにより取り出し、付属語などはキーワードから除外する。文書検索は以下のように行なう。
【００５６】
まず、以下の式で文書検索を行ない、上位k_{dr1}個の記事を取り出す。
【００５７】
【数１】

$$Score(d) = \sum_{\text{term } t} \left(\frac{tf(d,t)}{tf(d,t) + k_t \dfrac{length(d) + k_+}{\Delta + k_+}} \times log \frac{N}{df(t)} \right) \text{式(1)}$$

【００５８】
ただし、ｄは記事で、ｔは質問文から取り出したキーワードで、ｔｆ（ｄ，ｔ）は、記事ｄに出現するキーワードｔの頻度で、ｄｆ（ｔ）はキーワードｔが出現する頻度で、Ｎは記事の総数で、 length(d)は記事ｄの長さで、Δは記事長の平均である。k_tとk_+は実験で定める定数である。この式は、ロバートソンのＯｋａｐｉウェイティング（例えば、下記の文献（１）、文献（２）参照）の式に基づくもので、情報検索でよく用いられる式である（例えば、下記の文献（３）、文献（４）参照）。但し、質問応答では多くの種類のキーワードがマッチすることが重要なので、k_tの値としては大きな値を用いる。
【００５９】
文献（１）：S.E. Robertson and S.Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,(1994).
文献（２）：S.E. Robertson, S.Walker, S.Jones, M.M. HancockBeaulieu, and M.Gatford, Okapi at trec-3, TREC-3,(1994).
文献（３）：村田真樹，内元清貴，小作浩美，馬青，内山将夫，井佐原均，位置情報と

分野情報を用いた情報検索，言語処理学会誌，Ｖｏｌ．７，Ｎｏ．２（２０００）．

文献（４）：Masaki Murata, Qing Ma, and Hitoshi Isahara, High performance information retrieval using many characteristics and many techniques, Proceedings of the Third NTCIR Workshop (CLIR),(2002).

次に、以下の式で記事をリランキングし、上位k_{dr2}個の記事を取り出す。

【００６０】

【数２】

$$Score(d) = max_{t1 \in T} \sum_{t2 \in T3} w_{dr2}(t2) log \frac{N}{2dist(t1,t2) * df(t2)} \quad \text{式(2)}$$

$$T3 = \{t | t \in T, 2dist(t1,t) \frac{df(t)}{N} \leq 1\} \quad \text{式(3)}$$

【００６１】

ただし、Ｔはキーワードの集合で、ｄｉｓｔ（ｔ１，ｔ２）はキーワードｔ１とｔ２の間の距離で、便宜上ｔ１＝ｔ２のときｄｉｓｔ（ｔ１，ｔ２）＝０．５としている。w_{dr2}はｔ２の関数で実験により定められる。

【００６２】

一般には、質問応答システムでは質問文から取り出した複数のキーワードが近くに出現することを保証するために、記事を段落などの小さい単位に分割するが、非特許文献１のシステムでは、上記の、リランキングによりキーワードが近くにある場合に得点をあげる式を用いるので、記事を分割する必要がなく、記事をそのまま文書検索に使えるのである。この文書検索では、上位２０記事を取り出し、それを次の解の抽出で利用する。

【００６３】

（解の抽出）

文書検索で得た記事から、名詞、未知語連続を取り出し、それらを解の候補とする。それぞれの候補には、解の候補とキーワードの近さに基づく得点$Score_{near}$（ｃ）と解表現の意味制約を満足しているか否かに基づく$Score_{sem}$（ｃ）の二つの得点を与え、その合計点が最も大きい候補を解とする。

【００６４】

$Score_{near}$（ｃ）は、以下の式で与えられる。

【００６５】

【数３】

$$Score_{near}(c) = \sum_{t2 \in T3} w_{dr2}(t2) log \frac{N}{2dist(c,t2) * df(t2)} \quad \text{式(4)}$$

$$T3 = \{t | t \in T, 2dist(c,t) \frac{df(t)}{N} \leq 1\} \quad \text{式(5)}$$

【００６６】

ただし、ｃは解の候補であり、w_{dr2}は実験で定められる関数である。

【００６７】

解表現の意味制約に基づく得点$Score_{sem}$（ｃ）は、人手で作成した規則により与えられる。非特許文献１では、４５の規則を作成した。そのいくつかを以下に示す。
・推定した解表現（人名や地名など）と一致する候補に１０００を与える。解の候補が人名か地名かと特定する方法には、例えばＳＶＭに基づく固有表現抽出技術を利用する。固有表現抽出技術の例については、後述する。
・解表現が「国名」の場合に解の候補が国名のときに１０００を与える。
・質問文が「何＋名詞Ｘ」の場合、名詞Ｘを最後に持つ候補に１０００を与える。

【００６８】

非特許文献１における実験では、以下の得点加算法を利用している。

（１）オリジナル法

得点の加算を行わない方法。

（２）単純加算法

複数の記事から取り出した解の候補の得点を加算し、その得点をそのまま加算した合計得点に基づき解を出力する。

（３）逓減加点法

複数記事から取り出した候補の得点を加算する。この方法は、ｉ番目の候補の得点には$k^{(i-1)}$の値を乗じてから得点を加算する。すなわち、加算結果は、以下の式で表される。

【００６９】

【数４】

$$Score_{decreased} = \sum_{1 \leq i \leq n} k^{i-1} \, score_{original}(i) \qquad 式(6)$$

【００７０】

ただし、$Score_{decreased}$は、最終的な加算後の値の１０００より下の桁の数字で、$score_{original}$（ｉ）は、元の値の１０００より下の桁の数字である。ｎは１０００より上の桁で同じ数字を持つ複数の記事から得られた同じ解の候補の出現回数である。ｋは実験で定める定数である。

（４）融合法

この方法は、オリジナル法、単純加算法、逓減加点法の組み合わせである。この方法はまず学習データでこれらの方法のうちどの方法が最も良い精度を出すかを調べて、最も精度の高かった方法を利用して問題を解く。

【００７１】

この方法には融合による精度向上という効果と、公平な評価ができるという効果がある。

【００７２】

以下に、本発明の実施の形態について、図を用いて説明する。図１は、本発明の第１の実施の形態における質問応答装置の構成の一例を示す図である。第１の実施の形態では、例えば、第１のキーワード「日本」と第２のキーワード「面積」が入力されると、第１のキーワード「日本」に基づいて、第１のキーワードを、例えば「日本」、「アメリカ」、「ドイツ」という３つの第３のキーワードに増加させる。また、第２のキーワード「面積」に基づいて、第２のキーワードを、例えば「面積」、「人口」、「緯度」という３つの第４のキーワードに増加させる。そして、第３のキーワードと第４のキーワードとの組み合わせにより構成される、例えば「日本の面積は？」、「アメリカの人口は？」、「ドイツの緯度は？」・・・といった各質問に対する解答の候補を機械学習の手法を用いて求めて、解答として出力する。

【００７３】

質問応答装置１は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。質問応答装置１は、キーワード入力部１１

、キーワード増加部１２、質問作成部１３、解答候補抽出部１４、解答表出力部１５、キーワード抽出用データベース（ＤＢ）１６を備える。また、図中、１７は後述する解答候補抽出部１４による機械学習の結果（学習結果）が蓄積されている学習データベース（ＤＢ）である。

【００７４】

キーワード入力部１１には、キーワードが入力される。例えば、第１のキーワード「日本」と第２のキーワード「面積」が入力される。キーワード増加部１２は、後述するキーワード抽出技術を用いて、入力された各キーワードと同じ分野のキーワードをキーワード抽出用ＤＢ１６から抽出する。キーワードの抽出の結果、キーワードの総数が増加する。例えば、キーワード増加部１２は、入力された第１のキーワードに基づいて、第１のキーワードの数より多い第３のキーワードを出力する。また、例えば、キーワード増加部１２は、入力された第２のキーワードに基づいて、第２のキーワードの数より多い第４のキーワードを出力する。

【００７５】

質問作成部１３は、増加したキーワードである第３のキーワードと第４のキーワードとによって構成される質問を複数作成する。例えば、増加後の第３のキーワードの一つが「Ｘ」、第４のキーワードの一つが「Ｙ」とすると、所属の格助詞「の」を用いて、「ＸのＹは？」という質問を作成する。

【００７６】

解答候補抽出部１４は、後述する機械学習の手法によって、上記質問作成部１３によって作成された質問に対する解答の候補である解答候補を抽出する。解答表出力部１５は、抽出された各解答候補が質問と対応付けられた表を解答表として出力する。例えば、図２に示すような解答表を出力する。

【００７７】

図２に示す解答表の例では、例えば、「日本の面積は？」という質問に対応する解答として、解答表のデータ項目「日本」に対応する行とデータ項目「面積」と対応する列とが交差する枡目に「Ａ１」（km^2）が格納され、「アメリカの人口は？」という質問に対応する解答として、解答表のデータ項目「アメリカ」に対応する行とデータ項目「人口」と対応する列とが交差する枡目に「Ｂ２」（万人）が格納される。

【００７８】

本発明の実施の形態においては、抽出された解答候補を所定の単位（例えばkm^2）に換算した表現を解答表に格納してもよく、また、抽出された解答候補についての単位のまま解答表に格納してもよい。

【００７９】

もちろん、本発明において出力される解答表は、図２に示すものに限られるものではなく、例えば、「日本の面積は？→Ａ１（km^2）」、「アメリカの人口は？→Ｂ２（万人）」といった、各解答候補が矢印によって質問と対応付けられたデータが、解答表の各行のデータとして格納される形式の解答表を出力する構成を採ることもできる。

【００８０】

キーワード抽出用ＤＢ１６は、一定量の文書データを格納したデータベースである。キーワード抽出用ＤＢ１６は、例えば、新聞、雑誌、Ｗｅｂデータ（ネットワーク上のデータ）等から抽出したデータ（一定量の文書データ）を格納している。学習ＤＢ１７には、後述する学習結果が蓄積されている。例えば、『質問「日本の首都は？」で答え「東京」』という問題から抽出される素性の集合のときに、どのような解答（「正解」または「不正解」）になりやすいかが、学習結果として蓄積されている。

【００８１】

キーワード増加部１２は、パターン抽出部１２１とキーワード抽出部１２２とを備える。パターン抽出部１２１は、キーワード入力部１１に入力されたキーワードをキーワード抽出用ＤＢ１６で全文検索し、複数の入力キーワードの周辺に出現したパターンを抽出する。キーワード抽出部１２２は、パターン抽出部１２１で抽出したパターンをキーワード

抽出用ＤＢ１６で全文検索し、該パターンによって抽出される表現をキーワードとして出力する。

【００８２】

本発明の実施の形態においては、図１に示す構成から質問作成部１３を省略し、解答候補抽出部１４が、機械学習の手法を用いて、キーワード増加部１２によって出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。すなわち、解答候補抽出部１４は、予め用意された問題と、その問題に対する解答の組の多数のセットを用いて、どういう問題のときにどういう解答になるかを学習し、その学習結果に基づいて、キーワード増加部１２によって出力された第３のキーワードと第４のキーワードによって構成される質問に対する解答候補を抽出する構成を採ってもよい。

【００８３】

以下に、キーワード増加部１２によるキーワード抽出処理を説明する。パターン抽出部１２１は、入力された少数のキーワードをキーワード抽出用ＤＢ１６で全文検索し、該少数のキーワードの周辺に出現したパターンc_iを抽出する。キーワード抽出部１２２は、抽出したパターンc_iをキーワード抽出用ＤＢ１６で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅ（スコア；評価値）の値の大きい順にソートしてキーワードとして出力する。

【００８４】

本発明の実施の形態においては、キーワード抽出部１２２は、抽出した表現ｅｘｐについて、Ｓｃｏｒｅの値が大きいものから順に所定の個数取り出してキーワードとして出力する構成を採ってもよい。また、キーワード抽出部１２２は、抽出した表現ｅｘｐについて、Ｓｃｏｒｅの値が所定の閾値以上のものをキーワードとして出力する構成を採ってもよい。

【００８５】

（パターンの例の説明）

以下に、パターン抽出部１２１が抽出するパターンについて、該パターンが国名Ａである場合を例にとって説明する。

【００８６】

・入力キーワード：

日本
中国
朝鮮
タイ
韓国

・抽出パターンの例(1)：（両端とも利用、スピードは遅いが性能は良い）

日、Ａ軍
人のＡ人女性
日本はＡと
〔Ａ通信・
省。駐Ａ大使な

・抽出パターンの例(2)：（片方のみ利用、片方は平仮名文字、スピードは早い）

［..Ａ国］。

【００８７】

語。Ａ
［..Ａ国］側
［..Ａ国］伝来
Ａ語入力

ただし、［..Ａ..］は、それ自体が国名Ａにマッチすることを意味する。例えば［Ａ国］だとそのマッチした用語の最後が国であることを意味する。

【００８８】
（キーワード抽出の具体的な説明）
入力する少数のキーワードとして、例えば、評価データの代表形で毎日新聞での頻度の多い方から有名そうな用語を五つ選択するものとする。また、例えば、ＣＤ毎日新聞（コンパクトディスクに記録された毎日新聞）１９９１－２０００年度版をキーワード抽出用ＤＢ１６とする。抽出の手順例は以下のとおりである。
【００８９】
(1) 少数の複数のキーワードをキーワード抽出用ＤＢ１６で全文検索し、複数のキーワードの周辺に出現したパターンをc_iとして抽出する（キーワードの周辺に出現するパターンがそのキーワードだけ（一個）の場合は抽出しない）。（周辺に出現するパターンの定義は適宜行なう）。周辺に出現するパターンとして例えば、キーワードの前後（左右）３文字列を用いる場合は、前後それぞれ文字が１個、２個、３個の場合があるので、１個のキーワードで９通りのパターンができることになる。また、キーワード（自分自身）を含めたパターンとすることもできる。
【００９０】
(2) 次に抽出したパターンc_iをキーワード抽出用ＤＢ１６で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出する。
【００９１】
(3) 抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートして、キーワードとして出力する。
【００９２】
Ｓｃｏｒｅとして、以下のものがある。
【００９３】
・手法１（決定リスト法）
手法１は、抽出した表現ｅｘｐのＳｃｏｒｅとして、パターンc_iの中でp_iが最も大きかったパターンのp_iを使用する手法である。ここで、p_iはパターンc_iで抽出される表現ｅｘｐでの入力キーワードの割合（確からしさ、すなわち確信度となる）である。
【００９４】
例えば、パターンc_1についてキーワード抽出用ＤＢ１６で全文検索した結果、ｅｘｐ１、ｅｘｐ２、ｅｘｐ３、ｅｘｐ４、ｅｘｐ５までの５個のｅｘｐが抽出され、この５個のｅｘｐのうち、ｅｘｐ１～ｅｘｐ３までの３個が入力キーワードであった場合、p_1は３／５である。
【００９５】
【数５】

$$\mathrm{Score} = max_i p_i \qquad \text{式(7)}$$

【００９６】
・手法２（ベイズ法）
手法２は、抽出した表現ｅｘｐのＳｃｏｒｅとして、全てのパターンc_iのp_iを掛け合わせたものを使用する。
【００９７】
【数６】

$$\mathrm{Score} = \prod_i p_i \qquad \text{式(8)}$$

【００９８】
なお、実際には$p_i = 0$の可能性が大きいため、本発明の実施の形態では、上記式（８）に代えて、以下の式（９）

Π（（１－Δ）／Δ＊ｐ$_i$ ＋１）　　式（９）

を利用する構成をとることもできる。ここで、Δは微小値の定数であり、例えば、０．０００１を用いる。

【００９９】

例えば、Ｓｃｏｒｅを計算しているｅｘｐが、パターンｃ$_i$ についての検索処理によって取得できなかった場合は、ｐ$_i$ ＝０として、上記の式（９）を用いて計算する。

【０１００】

・手法３（類似度に基づく方法）

手法３は、抽出した表現ｅｘｐのＳｃｏｒｅとして、抽出されたパターンの個数（総数）を用いる。つまり、多くのパターンで抽出されたものほどＳｃｏｒｅを大きくする。

【０１０１】

【数７】

$$\mathrm{Score} = \sum_i 1 \qquad \text{式(10)}$$

【０１０２】

・手法４（下記研究（１）参照）

手法４は、抽出した表現ｅｘｐのＳｃｏｒｅとして、ｐ$_i$ の重みを加えた抽出されたパターンの個数を用いるものである。

【０１０３】

【数８】

$$\mathrm{Score} = \sum_i (1 + 0.01 p_i log(f_i)) \qquad \text{式(11)}$$

【０１０４】

ただし、ｆ$_i$ はパターンｃ$_i$ が出現した入力キーワードの個数である。

【０１０５】

研究（１）:Ellen Riloff and Rosie Jones "Learning dictionaries for information extraction by multi-level bootstrapping" Proceedings of AAAI-99,(1999) 。

【０１０６】

・手法５（下記文献（５）参照）

手法５は、抽出した表現ｅｘｐのＳｃｏｒｅとして、少なくとも一つは確からしくなる値を用いるものである。

【０１０７】

【数９】

$$\mathrm{Score} = 1 - \prod_i (1 - p_i) \qquad \text{式(12)}$$

【０１０８】

上記式（１２）は、確からしくない（１－ｐ$_i$ ）を掛け合わせることで一つも確からしくないことになり、そして、これを１から引くと少なくとも一つは確からしくなる。

【０１０９】

文献（５）：村田真樹，井佐原均 "同義テキストの照合に基づくパラフレーズに関する知識の自動獲得" 情報処理学会自然言語処理研究会 2001-NL-142,(2001) 。

【０１１０】

上記手法１、２、４、５では、Ｓｃｏｒｅが同じときは、手法３のＳｃｏｒｅでソート

し、手法３では手法５のＳｃｏｒｅでソートする。

【０１１１】

図３は、パターンとしてキーワードの左と先頭のいずれかを含む１～３文字と右側のそれの組み合わせを用いて行ったキーワードの抽出結果に対して、予め用意した所定の種類数の正解データを使って、適合率・再現率を求めた結果の一例を示す図である。ここで、正解データとしては、例えば、図４に示すようなデータ例を用意する（図４は、国名データの例を示しており、国名を国ごとに行に分けて格納し、行頭を代表形としてそれ以外は代表形の異表記として同じ行に格納している）。図４に示すデータ形式と同様のデータ形式を持つ正解データを、例えば、国名データの他に、衛星、祝日、太陽系惑星、世界遺産等に関するデータのように、多種類用意する。

【０１１２】

図３において、ＡＰは、情報検索（下記文献（６）参照）で用いるaverage precisionの平均であり、正解記事を上位から取ったたびに求めた適合率の平均である。本願の内容の場合は、正解キーワード分を上位から取ったたびに求めた適合率の平均（ただし、入力キーワードは正解キーワードから除く）である。

【０１１３】

文献（６）：村田真樹，馬青，内元清貴，小作浩美，内山将夫，井佐原均 "位置情報と分野情報を用いた情報検索" 言語処理学会誌，Vol.7,No.2,(2000)。

【０１１４】

ＲＰは、r-precision の平均であり、正解記事数分だけを検索した時に正解の記事が含まれている割合である。本願の内容の場合は、正解キーワード分だけを抽出した時に正解キーワードが含まれている割合である。なお、適合率は正解率と同じであり、正解キーワードが含まれる割合のことである。ＴＰは、上位５個での精度の平均である。

【０１１５】

（制約に基づく抽出方法の説明）

（ａ）字種とＫＲを利用する方法

図３に示す例で、抽出方法には、さらに字種とＫＲを利用する方法を用いた。ここで、字種とは、漢字、カタカナ、ひらがな、記号、数字などであり、例えば英語だと、アルファベット、数字、記号、単語の先頭が大文字かどうかなどである。

【０１１６】

字種を利用する方法では、入力した少数（この例では５個）のキーワードになかった字種を含む表現を抽出しない方法である。例えば、入力した５個のキーワードにひらがなが無かった場合は、ひらがなを含む表現を抽出しないようにするものである。

【０１１７】

ＫＲを利用する方法では、p_i を $p_i * f_i / n_i$ に置き換えた方法である。この方法の利点は、p_i が同じでも f_i / n_i の値により確信度を変えることができるものである。ただし、n_i は入力キーワードの個数で、手法３のときはＫＲの場合は１を f_i に置き換えた。なお、評価では抽出した結果でキーワードの異表記は除いた。また、字種による方法以外にも次のような方法もある。

【０１１８】

（ｂ）品詞に基づく方法

品詞に基づく方法では、例えば、入力表現に名詞しかない場合は出力時に名詞以外の表現を省く、また、入力表現に形容詞しかない場合は出力時に形容詞以外の表現を省くというものである。さらに、表現が複数の単語で構成されている場合は、末尾の単語（形態素）の品詞の情報を使うようにすることができる。

【０１１９】

（例による説明１）

入力キーワードとして次のものであった場合、

「楽しい」「哀しい」「嬉しい」「とても嬉しい」「とても哀しい」

抽出物として次のものが得られる場合、

「とても」「新しい」「美しい」「とても美しい」「とても難しい」
上記抽出物の表現中の末尾の単語の品詞を推定し、上記入力キーワードでは、末尾の単語の品詞は「形容詞」しかないので、抽出物の中で、末尾の単語の品詞が「形容詞」でない、副詞（「とても」）を除いて出力するようにする。
【０１２０】
（例による説明２）
入力キーワードとして次のものであった場合、
「楽しい」「歓喜」「悲痛」「悲しい」
上記入力キーワードでは、「形容詞」と「名詞」のように複数種類があった場合は、それらの品詞は出力し、それらの品詞以外の表現は出力しないようにする。
【０１２１】
なお、前述のような末尾の単語（形態素）の品詞の推定等の品詞情報を得るためには、次のような形態素解析システム（形態素解析手段）が必要になる。
【０１２２】
・形態素解析システムの説明
日本語を単語に分割するために、キーワード抽出部１２２で形態素解析システムを利用することが必要になる。ここではChaSenについて説明する（奈良先端大で開発されている形態素解析システム茶筌。http://chasen.aist-nara.ac.jp/index.html.jp で公開されている）。
【０１２３】
これは、日本語文を分割し、さらに、各単語の品詞も推定してくれる。例えば、「学校へ行く」を入力すると以下の結果を得ることができる。
【０１２４】
学校　ガッコウ　学校　名詞－一般
へ　へ　へ　助詞－格助詞－一般
行く　イク　行く　動詞－自立　五段・カ行促音便　基本形
ＥＯＳ
このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。
【０１２５】
（ｃ）共通部分文字列に基づく方法
例えば、入力表現がすべて同じ「しい」という共通末尾表現を持っている場合、出力時に「しい」を持たない表現を省くものである。なお、これは末尾だけでなく、先頭の文字列でも同様にできる。
【０１２６】
（例による説明）
入力キーワードとして次のものであった場合、
「悲しい」「楽しい」「嬉しい」
抽出されるものが次の場合、
「歓喜」「悲痛」「美しい」「新しい」
上記入力キーワードの共通部分文字列が「しい」なので、「しい」を持たない「歓喜」と「悲痛」を削除して出力するものである。
【０１２７】
（ｄ）ユーザによる制約の指定
上記では、入力表現から自動で制約を得る方法を説明したが、この制約はユーザにさせることもできる。例えば、ユーザが「漢字のみ」というオプションを選択すると出力では漢字以外の字種を用いた表現を出力しないことができる。また、ユーザが末尾は「しい」というオプションを選択すると出力では「しい」を末尾に持たない表現を出力しないようにすることができる。さらに、ユーザが品詞は名詞というオプションを選択すると出力では名詞以外の表現を出力しないようにする。

【０１２８】
次に、質問作成部１３が作成した質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問（「ＸのＹは？」）に対する解答候補を抽出する処理について説明する。解答候補抽出部１４は、機械学習の手法を用いて解答候補を抽出する。
（機械学習の手法）
機械学習の手法は、問題－解答の組のセットを多く用意し、それで学習を行ない、どういう問題のときにどういう解答になるかを学習し、その学習結果を利用して、新しい問題のときも解答を推測できるようにする方法である（例えば、下記の文献（７）参照）。
【０１２９】
文献（７）：村田真樹，機械学習に基づく言語処理，龍谷大学理工学部．招待講演．2004．http://www2.nict.go.jp/jt/a132/members/murata/ps/rk1-siryou.pdf
どういう問題のときに、という、問題の状況を機械に伝える際に、素性（解析に用いる情報で問題を構成する各要素）というものが必要になる。問題を素性によって表現するのである。例えば、日本語文末表現の時制の推定の問題において、問題：「彼が話す。」－－－解答「現在」が与えられた場合に、素性の一例は、「彼が話す。」「が話す。」「話す。」「す」「。」となる。
【０１３０】
すなわち、機械学習の手法は、素性の集合－解答の組のセットを多く用意し、それで学習を行ない、どういう素性の集合のときにどういう解答になるかを学習し、その学習結果を利用して、新しい問題のときもその問題から素性の集合を取り出し、その素性の場合の解答を推測する方法である。
【０１３１】
まず、機械学習の手法一般についての説明をする。機械学習の手法としては、一般に、ｋ近傍法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などの手法を用いる。
【０１３２】
ｋ近傍法は、最も類似する一つの事例のかわりに、最も類似するｋ個の事例を用いて、このｋ個の事例での多数決によって分類先（解）を求める手法である。ｋは、あらかじめ定める整数の数字であって、一般的に、１から９の間の奇数を用いる。
【０１３３】
シンプルベイズ法は、ベイズの定理にもとづいて各分類になる確率を推定し、その確率値が最も大きい分類を求める分類先とする方法である。
【０１３４】
シンプルベイズ法において、文脈ｂで分類ａを出力する確率は、以下の式（１３）で与えられる。
【０１３５】
【数１０】

$$p(a|b)=\frac{p(a)}{p(b)}p(b|a) \qquad \text{式(13)}$$

$$\cong\frac{\tilde{p}(a)}{p(b)}\prod_i\tilde{p}(f_i|a) \qquad \text{式(14)}$$

【０１３６】
ただし、ここで文脈ｂは、あらかじめ設定しておいた素性f_j（∈Ｆ，１≦ｊ≦ｋ）の

集合である。ｐ（ｂ）は、文脈ｂの出現確率である。ここで、分類ａに非依存であって定数のために計算しない。Ｐ（ａ）（ここでＰはｐの上部にチルダ）とＰ（f_i｜ａ）は、それぞれ教師データから推定された確率であって、分類ａの出現確率、分類ａのときに素性f_iを持つ確率を意味する。Ｐ（f_i｜ａ）として最尤推定を行って求めた値を用いると、しばしば値がゼロとなり、式（１４）の値がゼロで分類先を決定することが困難な場合が生じる。そのため、スームージングを行う。ここでは、以下の式（１５）を用いてスームージングを行ったものを用いる。

【０１３７】

【数１１】

$$p(f_i|a)=\frac{freq(f_i,a)+0.01*freq(a)}{freq(a)+0.01*freq(a)} \qquad \text{式(15)}$$

【０１３８】

ただし、ｆｒｅｑ（f_i，ａ）は、素性f_iを持ちかつ分類がａである事例の個数、ｆｒｅｑ（ａ）は、分類がａである事例の個数を意味する。

【０１３９】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力が与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

【０１４０】

決定リスト方法では、あらかじめ設定しておいた素性f_j（∈Ｆ，１≦ｊ≦ｋ）のうち、いずれか一つの素性のみを文脈として各分類の確率値を求める。ある文脈ｂで分類ａを出力する確率は以下の式によって与えられる。

【０１４１】

ｐ（ａ｜ｂ）＝ｐ（ａ｜f_{max}）　　式（１６）

ただし、f_{max}は以下の式によって与えられる。

【０１４２】

【数１２】

$$f_{\max}=\arg\max_{f_j\in F}\ \max_{a_i\in A}\ \tilde{p}(a_i|f_j) \qquad \text{式(17)}$$

【０１４３】

また、Ｐ（a_i｜f_j）（ここでＰはｐの上部にチルダ）は、素性f_jを文脈に持つ場合の分類a_iの出現の割合である。

【０１４４】

最大エントロピー法は、あらかじめ設定しておいた素性f_j（１≦ｊ≦ｋ）の集合をＦとするとき、以下所定の条件式（式（１８））を満足しながらエントロピーを意味する式（１９）を最大にするときの確率分布ｐ（ａ，ｂ）を求め、その確率分布にしたがって求まる各分類の確率のうち、最も大きい確率値を持つ分類を求める分類先とする方法である。

【０１４５】

【数１３】

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b)=\sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad \text{式(18)}$$

$$for\ \forall f_j(1\le j\le k)$$

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)\log(p(a,b)) \quad \text{式(19)}$$

【０１４６】

ただし、Ａ、Ｂは分類と文脈の集合を意味し、g_j（ａ，ｂ）は文脈ｂに素性f_jがあって、なおかつ分類がａの場合１となり、それ以外で０となる関数を意味する。また、Ｐ（a_i｜f_j）（ここでＰはｐの上部にチルダ）は、既知データでの（ａ，ｂ）の出現の割合を意味する。

【０１４７】

式（１８）は、確率ｐと出力と素性の組の出現を意味する関数ｇをかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化(確率分布の平滑化) を行なって、出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については、以下の文献（８）および文献（９）に記載されている。

【０１４８】

文献（８）：Eric Sven Ristad, Maximum Entropy Modeling for Natural Language,(ACL/EACL Tutorial Program, Madrid, 1997

文献（９）：Eric Sven Ristad, Maximum Entropy Modeling Toolkit, Release 1.6beta, (http://www.mnemonic.com/software/memt,1998))

サポートベクトルマシン法は、空間を超平面で分割することにより、二つの分類からなるデータを分類する手法である。

【０１４９】

図２７にサポートベクトルマシン法のマージン最大化の概念を示す。図２７において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図２７（Ａ）は、正例と負例の間隔が狭い場合（スモールマージン）の概念図、図２７（Ｂ）は、正例と負例の間隔が広い場合（ラージマージン）の概念図である。

【０１５０】

このとき、二つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔（マージン）が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図２７（Ｂ）に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【０１５１】

基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線形にする拡張（カーネル関数の導入）がなされたものが用いられる。

【０１５２】
この拡張された方法は、以下の識別関数を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。
【０１５３】
【数１４】

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i, x)+b\right) \qquad \text{式(20)}$$

$$b = -\frac{\max_{i,y_i=-1} b_i + \min_{i,y_i=1} b_i}{2}$$

$$b_i = \sum_{j=1}^{l}\alpha_j y_j K(x_j, x_i)$$

【０１５４】
ただし、ｘは識別したい事例の文脈（素性の集合）を、x_i と y_j （ｉ＝１，…，ｌ，$y_j \in \{1, -1\}$）は学習データの文脈と分類先を意味し、関数ｓｇｎは、

ｓｇｎ（ｘ）＝１（ｘ≧０）
　　－１（otherwise ）

であり、また、各α_iは式（２２）と式（２３）の制約のもと式（２１）を最大にする場合のものである。
【０１５５】
【数１５】

$$L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \qquad \text{式(21)}$$

$$0 \le \alpha_i \le C \quad (i = 1,...,l) \qquad \text{式(22)}$$

$$\sum_{j=1}^{l}\alpha_i y_i = 0 \qquad \text{式(23)}$$

【０１５６】
また、関数Ｋはカーネル関数と呼ばれ、様々なものが用いられるが、本形態では以下の多項式のものを用いる。
【０１５７】
Ｋ（ｘ，ｙ）＝（ｘ・ｙ＋１）ｄ　　式（２４）
Ｃ、ｄは実験的に設定される定数である。後述する具体例ではＣはすべての処理を通して１に固定した。また、ｄは、１と２の二種類を試している。ここで、$\alpha_i > 0$となるx_iは、サポートベクトルと呼ばれ、通常、式（２０）の和をとっている部分は、この事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。
【０１５８】

なお、拡張されたサポートベクトルマシン法の詳細については、以下の文献（１０）および文献（１１）に記載されている。

【０１５９】

文献（１０）：Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods,(Cambridge University Press,2000)

文献（１１）：Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aist-nara.ac.jp/taku-ku//software/Tiny SVM/index.html,2000)

サポートベクトルマシン法は、分類の数が２個のデータを扱うものである。したがって、分類の数が３個以上の事例を扱う場合には、通常、これにペアワイズ法またはワンＶＳレスト法などの手法を組み合わせて用いることになる。

【０１６０】

ペアワイズ法は、ｎ個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア（ｎ（ｎ－１）／２個）を生成し、各ペアごとにどちらがよいかを二値分類器、すなわちサポートベクトルマシン法処理モジュールで求めて、最終的に、ｎ（ｎ－１）／２個の二値分類による分類先の多数決によって、分類先を求める方法である。

【０１６１】

ワンＶＳレスト法は、例えば、ａ、ｂ、ｃという三つの分類先があるときは、分類先ａとその他、分類先ｂとその他、分類先ｃとその他、という三つの組を生成し、それぞれの組についてサポートベクトルマシン法で学習処理する。そして、学習結果による推定処理において、その三つの組のサポートベクトルマシンの学習結果を利用する。推定すべき二項関係の候補が、その三つのサポートベクトルマシンではどのように推定されるかを見て、その三つのサポートベクトルマシンのうち、その他でないほうの分類先であって、かつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を求める解とする方法である。例えば、ある候補が、「分類先ａとその他」の組の学習処理で作成したサポートベクトルマシンにおいて分離平面から最も離れた場合には、その候補の分類先はa と推定する。

【０１６２】

次に、本発明の実施の形態における具体的な解答候補の抽出手法を説明する。

（解答候補の抽出手法１）

＜問題の構成＞

予め、

問題『質問「Ｘ１のＹ１は？」で答え「Ｚ１」』－－－解答「正解」

問題『質問「Ｘ２のＹ２は？」で答え「Ｚ２」』－－－解答「正解」

問題『質問「Ｘ３のＹ３は？」で答え「Ｚ３」』－－－解答「不正解」

という、問題と解答の対を多数作成する。

【０１６３】

また、例えば、上記の問題を表現する、以下のような素性を用意する。

・Ｘｉ，Ｙｉ，Ｚｉの単語自体

・Ｘｉ，Ｙｉ，Ｚｉの単語の意味クラス

・Ｘｉ，Ｙｉで検索した記事数

・Ｘｉ，Ｙｉで検索した記事にＺｉが存在する記事数

・Ｘｉ，Ｙｉが近接して（ある単語数の範囲内に）出現した記事数

・Ｘｉ，Ｙｉ，Ｚｉが近接して（ある単語数の範囲内に）出現した記事数

・Ｘｉ，Ｙｉで検索した記事に最も多く出現した単語とＺｉが一致するかどうか

・Ｘｉ，Ｙｉで検索した記事にｊ番目に多く出現した単語とＺｉが一致するかどうか

・Ｘｉ，Ｙｉをキーワードとして、例えば、解答候補抽出部１４が、新聞記事データ・百科事典データなどの文書データ群（図示を省略）から解答の書いてありそうな記事群を取り出し、その取り出した記事群の言語表現を解答の候補として取り出し、取り出された解答の候補を、優先順序（例えば、$Score_{near}$（ｃ））で並び替えた場合に、その順序

の最も高い候補とＺｉが一致したかどうか、また、その順序のｊ番目の候補とＺｉが一致したかどうか

上記の処理によって、素性の集合と解答の組の多数のセットが用意される。

【０１６４】

ここで、優先順序として用いる$Score_{near}$（ｃ）については、前述の非特許文献１に記述されており、解答の候補とキーワードの近さに基づく得点を示している。

【０１６５】

次に、意味クラスを説明する。一般に、各単語がどういう意味クラスを持つかを記述した表があり、その表を使えば、単語の意味クラスを求めることができる。例えば分類語彙表がある。分類語彙表では単語は分類番号と呼ばれる１０桁の数字で表現され、この数字の良く似ている単語ほど良く似た単語となる。例えば、この数字の最初の３桁や５桁を単語の意味クラスとして利用する。例えば、「村人」の分類番号は１２３０１０２０５０であり、これは１２３（人種、国民、社会階層などの意味クラス）、１２３０１（国民、住民などの意味クラス）に属する単語であることが示される。

【０１６６】

問題構成と素性の定義をすれば、あとは機械学習の手法で扱える。すなわち、解答候補抽出部１４は、用意された素性と解答の組の多数のセットを用いて、どういう素性の集合のときにどういう解答になるかを学習し、その学習結果を利用して、新たな問題についての素性の集合の場合に推測される解答を、解答候補として抽出する。

【０１６７】

＜問題や素性の具体例＞

問題の具体例：

問題『質問「日本の首都は？」で答え「東京」』－－－解答「正解」

問題『質問「日本の首都は？」で答え「大阪」』－－－解答「不正解」

問題『質問「日本の首都は？」で答え「パン」』－－－解答「不正解」

素性の具体例：

問題『質問「日本の首都は？」で答え「東京」』－－－解答「正解」の場合

・Ｘｉの単語自体：日本

・Ｙｉの単語自体：首都

・Ｚｉの単語自体：東京

・Ｘｉの意味クラス：１２５９０（地名のクラス）

・Ｙｉの意味クラス：１２５４０（都市集落のクラス）

・Ｚｉの意味クラス：１２５９０（地名のクラス）

（意味クラスとして分類語彙表の最初の５桁を利用する。）

・Ｘｉ，Ｙｉで検索した記事数：日本と首都を含む記事数。例えば１０００

・Ｘｉ，Ｙｉで検索した記事にＺｉが存在する記事数：日本と首都と東京を含む記事数。例えば１００

・Ｘｉ，Ｙｉが近接して（ある単語数の範囲内に）出現した記事数：例えば、日本と首都が１０単語以内にある記事数。例えば５００

・Ｘｉ，Ｙｉ，Ｚｉが近接して（ある単語数の範囲内に）出現した記事数：例えば、日本と首都と東京が１０単語以内にある記事数。例えば５０

・Ｘｉ，Ｙｉで検索した記事に最も多く出現した単語とＺｉが一致するかどうか：例えば、ここでは特に単語は名詞にしぼり、名詞としては、「こと」が最も頻度が多かったとすると、「こと」と「東京」が一致しないのでこの素性は「いいえ」となる。

・Ｘｉ，Ｙｉで検索した記事にｊ番目に多く出現した単語とＺｉが一致するかどうか：例えば、ここでは特に単語は名詞にしぼり、名詞としては、「東京」が二番目に頻度が多かったとすると、ｊ＝２の場合の素性は「はい」となる。

・Ｘｉ，Ｙｉをキーワードとして、例えば、解答候補抽出部１４が、新聞記事データ・百科事典データなどの文書データ群から解答の書いてありそうな記事群を取り出し、その取り出した記事群の言語表現を解答の候補を取り出し、取り出された解答の候補を、優先順

序（例えば、$Score_{near}$（ｃ））で並び替えた場合に、その順序の最も高い候補とＺｉが一致したかどうか、また、その順序のｊ番目の候補とＺｉが一致したかどうか
例えば、１番目の候補が「こと」、２番目の候補が「東京」の場合は、「その順序の最も高い候補とＺｉが一致したかどうか」は「いいえ」になり、「その順序のｊ番目の候補とＺｉが一致したかどうか」は、ｊ＝２のとき「はい」になる。

【０１６８】

より多くの事例で学習すると、例えば、解答候補抽出部１４は、
Ｙｉの単語自体：首都
Ｚｉの意味クラス：１２５９０（地名のクラス）
で、「その順序の最も高い候補とＺｉが一致したかどうか」は「はい」
または「その順序のｊ番目の候補とＺｉが一致したかどうか」のｊ＝２のときが「はい」になれば、
解答「正解」
となるように学習し、
Ｙｉの単語自体：首都
Ｚｉの意味クラス：１２５９０（地名のクラス）以外なら、
解答「不正解」
Ｙｉの単語自体：首都
「その順序の最も高い候補とＺｉが一致したかどうか」は「いいえ」かつ、
「その順序のｊ番目の候補とＺｉが一致したかどうか」のｊ＝２から１０全てで「いいえ」ならば、
解答「不正解」
といったことを学習する。

【０１６９】

学習結果は、解答候補抽出部１４によって、学習ＤＢ１７中に蓄積される。そして、解答候補抽出部１４は、学習ＤＢ１７中の学習結果情報を使い、例えば、新しい問題（すなわち、キーワード増加部１２によって出力された出力キーワードによって構成される質問）：
質問「フランスの首都は？」－答え「パリ」については、
Ｙｉの単語自体：首都
Ｚｉの意味クラス：１２５９０（地名のクラス）で、
「その順序の最も高い候補とＺｉが一致したかどうか」は「はい」または、
「その順序のｊ番目の候補とＺｉが一致したかどうか」のｊ＝２のときが「はい」
なので、「正解」と判断する。

【０１７０】

また、新しい問題：
質問「フランスの首都は？」－答え「信号」については、上記学習結果を用いて、
Ｙｉの単語自体：首都
Ｚｉの意味クラス：１２５９０（地名のクラス）以外
なので、「不正解」と判断する。

【０１７１】

ここで、機械学習の手法によって、解答を求めるだけでなく、その解答がどのくらい正解になりやすいかの度合い、どのくらい不正解になりやすいかの度合いも同時に求めることができる。

【０１７２】

すなわち、解答候補抽出部１４は、用意した素性の集合と解答の組の多数のセットを用いて、まず、どのような素性の集合のときにどのような解答（正解または不正解）となるかということを機械学習し、どのような素性の集合のときにどのような解答となるかということを示す情報を、学習結果情報として学習ＤＢ１７に格納する。そして、解答候補抽出部１４は、新たな問題（キーワード増加部１２によって出力された出力キーワードによ

って構成される質問）を用いて、新たに解答候補抽出部１４が作成した問題）から素性の集合を抽出し、抽出された素性の集合の場合にどのような解答になりやすいか、すなわち、「正解となりやすい」かの度合いを、学習ＤＢ１７に格納された学習結果情報に基づいて求める。

【０１７３】

そして、解答候補抽出部１４は、例えば、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における、「答え」を、解答候補として解答表出力部１５に対して出力する。解答表出力部１５は、各解答候補が質問と対応付けられた表を解答表として出力する。本発明の実施の形態においては、例えば、「正解となりやすい」かの度合いの大きい順に所定の個数の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが所定の閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが最も大きい問題についての当該度合いの所定の割合（例えば、９０％等）を閾値とし、「正解となりやすい」かの度合いがこの閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。

【０１７４】

上記の、「正解となりやすい」かの度合いの求め方は、解答候補抽出部１４が機械学習の手法として用いる様々な方法によって異なる。

【０１７５】

例えば、本発明の実施の形態において、解答候補抽出部１４が、機械学習の手法としてｋ近傍法を用いる場合、解答候補抽出部１４は、上記用意した素性の集合と解答の組の多数のセットを用いて、素性の集合のうち重複する素性の割合（同じ素性をいくつ持っているかの割合）に基づく問題同士の類似度を定義する。そして、解答候補抽出部１４は、定義した類似度と問題（と解答）とを、学習結果情報として、学習ＤＢ１７に格納しておく。

【０１７６】

そして、解答候補抽出部１４は、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問を用いて新たに問題を作成すると、学習ＤＢ１７内に格納された類似度と問題を参照して、新たに作成された問題との類似度が高い順にｋ個の問題（と解答）を学習ＤＢ１７に格納された問題（と解答）から選択し、選択したｋ個の問題での多数決によって決まった分類先（正解または不正解）を、新たに作成された問題に対する解答とする。ｋ近傍法を用いる場合、「正解となりやすい」かの度合いは、上記選択されたｋ個の問題での多数決の票数、すなわち、「正解」という分類が獲得した票数となる。

【０１７７】

解答候補抽出部１４は、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における、「答え」を、解答候補として解答表出力部１５に対して出力する。本発明の実施の形態においては、例えば、「正解となりやすい」かの度合いの大きい順に所定の個数の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが所定の閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが最も大きい問題についての当該度合いの所定の割合（例えば、９０％等）を閾値とし、「正解となりやすい」かの度合いがこの閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。

【０１７８】

また、例えば、本発明の実施の形態において、解答候補抽出部１４が、機械学習の手法としてシンプルベイズ法を用いる場合、例えば、解答候補抽出部１４は、上記用意した素性の集合と解答の組の多数のセットを学習結果情報として学習ＤＢ１７に格納しておく。

【０１７９】

解答候補抽出部１４は、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問を用いて新たに問題を作成すると、新たに作成した問題から素性の集合を抽出する。そして、解答候補抽出部１４は、学習ＤＢ１７内に格納された解答と素性の集合とのセットをもとに、ベイズの定理に基づいて、新たに作成された問題から抽出した素性の集合の場合の各分類になる確率を算出して、その確率の値が最も大きい分類を、その問題に対する解答とする。シンプルベイズ法を用いる場合、「正解となりやすい」かの度合いは、「正解」という分類になる確率となる。

【０１８０】

解答候補抽出部１４は、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における、「答え」を、解答候補として解答表出力部１５に対して出力する。本発明の実施の形態においては、例えば、「正解となりやすい」かの度合いの大きい順に所定の個数の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが所定の閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが最も大きい問題についての当該度合いの所定の割合（例えば、９０％等）を閾値とし、「正解となりやすい」かの度合いがこの閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。

【０１８１】

また、例えば、本発明の実施の形態において、解答候補抽出部１４が、機械学習の手法として決定リスト法を用いる場合、例えば、解答候補抽出部１４は、予め用意した問題についての素性と分類先との規則を所定の優先順序で並べたリストを学習ＤＢ１７内に格納する。解答候補抽出部１４は、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問を用いて新たに問題を作成すると、新たに作成した問題から素性の集合を抽出する。

【０１８２】

そして、解答候補抽出部１４は、学習ＤＢ１７内に格納されたリストの優先順位の高い順に、上記新たに作成した問題から抽出された素性と規則の素性とを比較し、素性が一致した規則の分類先をその問題に対する解答とする。決定リスト法を用いる場合、「正解となりやすい」かの度合いは、所定の優先順位またはそれに相当する数値、尺度となる。

【０１８３】

解答候補抽出部１４は、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における、「答え」を、解答候補として解答表出力部１５に対して出力する。本発明の実施の形態においては、例えば、「正解となりやすい」かの度合いの大きい順に所定の個数の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが所定の閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが最も大きい問題についての当該度合いの所定の割合（例えば、９０％等）を閾値とし、「正解となりやすい」かの度合いがこの閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。

【０１８４】

また、例えば、本発明の実施の形態において、解答候補抽出部１４が、機械学習の手法として最大エントロピー法を用いる場合、例えば、解答候補抽出部１４は、予め用意した問題の解答となりうる分類を特定し、所定の条件式を満足しかつエントロピーを示す式を最大にするときの素性の集合と解答となりうる分類の二項からなる確率分布を求めて、学習ＤＢ１７内に格納する。

【０１８５】

そして、解答候補抽出部１４は、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によっ

て出力された出力キーワードによって構成される質問を用いて新たに問題を作成すると、学習ＤＢ１７内に格納された確率分布を利用して、新たな問題の素性の集合についてその解答となりうる分類の確率を求めて、最も大きい確率値を持つ解答となりうる分類を特定し、その特定した分類をその問題に対する解答とする。

【０１８６】

すなわち、最大エントロピー法を用いる場合、「正解となりやすい」かの度合いは、「正解」という分類になる確率となる。

【０１８７】

解答候補抽出部１４は、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における、「答え」を、解答候補として解答表出力部１５に対して出力する。本発明の実施の形態においては、例えば、「正解となりやすい」かの度合いの大きい順に所定の個数の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが所定の閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが最も大きい問題についての当該度合いの所定の割合（例えば、９０％等）を閾値とし、「正解となりやすい」かの度合いがこの閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。

【０１８８】

また、例えば、本発明の実施の形態において、解答候補抽出部１４が、機械学習の手法としてサポートベクトルマシン法を用いる場合、例えば、解答候補抽出部１４は、予め用意した問題の解答となりうる分類を特定し、分類を正例と負例に分割して、カーネル関数を用いた所定の実行関数に従って問題の素性の集合を次元とする空間上で、その問題の正例と負例の間隔を最大にし、かつ正例と負例を超平面で分割する超平面を求めて学習ＤＢ１７内に格納する。

【０１８９】

そして、解答候補抽出部１４は、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問を用いて新たに問題を作成すると、学習ＤＢ１７内の超平面を利用して、新たな問題の素性の集合が超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果に基づいて定まる分類を、その問題に対する解答とする。

【０１９０】

すなわち、サポートベクトルマシン法を用いる場合、「正解となりやすい」かの度合いは、分離平面からの正例の空間への距離の大きさとなる。より詳しくは、解答が正解である問題を正例、解答が不正解である問題を負例とする場合に、分離平面に対して正例側の空間に位置する問題が、解答が正解である問題と判断され、分離平面からの距離が大きい問題ほど「正解となりやすい」かの度合いが大きくなる。

【０１９１】

解答候補抽出部１４は、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における、「答え」を、解答候補として解答表出力部１５に対して出力する。本発明の実施の形態においては、例えば、「正解となりやすい」かの度合いの大きい順に所定の個数の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが所定の閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。また、例えば、「正解となりやすい」かの度合いが最も大きい問題についての当該度合いの所定の割合（例えば、９０％等）を閾値とし、「正解となりやすい」かの度合いがこの閾値以上の問題を選択し、選択した問題における「答え」を解答候補としてもよい。

【０１９２】

＜具体例＞

例えば、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成

部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問「フランスの首都は？」を解く場合について説明する。

【０１９３】

まず、解答候補抽出部１４が、出力キーワード「フランス」、「首都」を含む文書を、新聞記事データ・百科事典データなどの文書データ群から取得する。質問「フランスの首都は？」からの、キーワード「フランス」、「首都」の取り出しには、形態素解析技術などを使う。質問応答装置１が質問作成部１３を省略し、解答候補抽出部１４がキーワード増加部１２によって出力された出力キーワードにより構成される質問に対する解答候補を抽出する構成を採る場合には、上記の質問「フランスの首都は？」からキーワード「フランス」、「首都」を形態素解析技術を用いて取り出す必要はなく、解答候補抽出部１４は、キーワード増加部１２によって出力された出力キーワード「フランス」、「首都」をそのまま用いて、それらを含む文書を上記文書データ群から取得する。

【０１９４】

解答候補抽出部１４は、キーワード「フランス」、「首都」を含む文書中の言語表現を、質問の答えの表現の候補として取り出す。この表現の取り出しには、例えば、前述の非特許文献１に記載された解の抽出の処理を用いる。

【０１９５】

取り出された答えの表現の候補を、例えば、$Score_{near}$（ｃ）の値の大きい順に並び替え、その値の上位何個かの候補を取り出し、その候補（候補１，候補２，・・・）について、

問題『質問「フランスの首都は？」で答え「候補１」』
問題『質問「フランスの首都は？」で答え「候補２」』
・・・

を作成する。ここで、「候補１」、「候補２」は、上記の質問の「答え」の表現の候補を示している。

【０１９６】

作成された問題（質問－答えの対）について、前述した機械学習の手法を適用し、「正解となりやすい」かの度合いが最も大きいときの、問題（質問－答えの対）における「答え」を、解答候補として、解答表出力部１５に対して出力する。

【０１９７】

解答表出力部１５は、解答表において、質問「フランスの首都は？」に対する解答が格納される枡目（例えば、データ項目「フランス」に対応する行とデータ項目「首都」に対応する列とが交差する枡目）に、対応する解答候補を格納する。

【０１９８】

（解答候補の抽出手法２）

＜問題の構成＞

解答候補抽出部１４は、
問題『質問「Ｘ１のＹ１は？」』－－－解答「地名」
問題『質問「Ｘ２のＹ２は？」』－－－解答「地名」
問題『質問「Ｘ３のＹ３は？」』－－－解答「人名」
問題『質問「Ｘ４のＹ４は？」』－－－解答「数値」
・・・
という、問題と解答の対を多数作成する。
素性としては、
Ｘｉ，Ｙｉの単語自体
Ｘｉ，Ｙｉの単語の意味クラス
などが考えられる。
問題構成と素性の定義をすれば、あとは機械学習の手法で扱える。

【０１９９】

＜問題や素性の具体例＞

問題の具体例：
問題『質問「日本の首都は？」』－－－解答「地名」
問題『質問「日本の首相は？」』－－－解答「人名」
問題『質問「日本の面積は？」』－－－解答「数値」
素性の具体例：
問題『質問「日本の首都は？」』－－－解答「地名」の場合、
・Ｘｉの単語自体：日本
・Ｙｉの単語自体：首都
・Ｘｉの意味クラス：１２５９０（地名のクラス）
・Ｙｉの意味クラス：１２５４０（都市集落のクラス）
（意味クラスとして分類語彙表の最初の５桁を利用）
もっと多くの事例で学習すると、例えば、
Ｙｉの単語自体：首都
だと、
解答「地名」
となるように学習し、
Ｙｉの単語自体：首相
だと、
解答「人名」、
Ｙｉの単語自体：面積
だと、
解答「数値」、
といったことを、解答候補抽出部１４が学習し、その学習結果を学習ＤＢ１７内に蓄積する。

【０２００】

そして、解答候補抽出部１４は、学習ＤＢ１７内に蓄積された学習結果を用いて、解答を判断する。
例えば、新しい問題：
『質問「フランスの首都は？」』についての解答は、
Ｙｉの単語自体：首都
なので、「地名」と判断する。

【０２０１】

＜具体例＞
例えば、質問作成部１３によって作成された質問、または、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワードによって構成される質問「フランスの首都は？」を解く場合について説明する。

【０２０２】

まず、解答候補抽出部１４は、前述した機械学習の手法を利用して、
問題「フランスの首都は？」について、
解答が「地名」であるという結果を取得する。

【０２０３】

取得された解答「地名」を、解答表の枡目に格納する解答候補を抽出する際の解答タイプとして利用する。

【０２０４】

すなわち、解答候補抽出部１４は、新聞記事データ・百科事典データなどの文書データ群から、質問作成部１３が作成した質問「フランスの首都は？」を構成するキーワード（「フランス」、「首都」）を含む文書を取り出し、取り出された文書に含まれる言語表現のうち、上記解答タイプに適合するものを解答候補として解答表出力部１５に対して出力する。解答候補抽出部１４は、質問応答装置１が質問作成部１３を備えない構成を採るときはキーワード増加部１２によって出力された出力キーワード（「フランス」、「首都」

）を含む文書を上記文書データ群から取り出し、取り出された文書に含まれる言語表現のうち、上記解答タイプに適合するものを、出力キーワード「フランス」と「首都」とによって構成される質問「フランスの首都は？」に対する解答候補として解答表出力部１５に対して出力する。

【０２０５】

解答表出力部１５は、解答表において、質問「フランスの首都は？」に対する解答が格納される枡目（例えば、データ項目「フランス」に対応する行とデータ項目「首都」に対応する列とが交差する枡目）に、対応する解答候補を格納する。

【０２０６】

なお、本発明においては、例えば、「ＸのＹは？」という質問に対する解答候補を抽出する際に、例えば、解答候補抽出部１４が、機械学習の手法を用いるのではなく、新聞記事データ・百科事典データなどの大量の文書データ群（図示を省略）からキーワード「Ｘ」とキーワード「Ｙ」を含む記事群を取り出し、その取り出した記事群の言語表現のうち、上記文書データ群中に出現する頻度が所定の閾値以上のものを解答候補として出力する構成を採ることもできる。また、本発明の実施の形態においては、上記取り出した記事群の言語表現について、上記文書データ群中に出現する頻度の高い順に所定の個数取り出して、解答候補として出力する構成を採ることもできる。

【０２０７】

ここで、上記の解答候補抽出部１４による、解答タイプを用いた解答候補の出力の際には、非特許文献１の説明において述べた固有表現抽出技術を用いる。固有表現とは、人名、地名、組織名などの固有名詞、金額などの数値表現といった、特定の事物・数量を意味する言語表現のことで、固有表現抽出とは、そういった固有表現を文章中から計算機で自動で抽出する技術である．例えば、「日本の首相は小泉純一郎である」という文に対して固有表現抽出を行なうと、固有表現の「日本」と「小泉純一郎」が地名、人名として、抽出される。本発明の実施の形態においては、解答候補抽出部１４が、抽出された固有表現が上記解答タイプに適合するかを判断し、適合する固有表現を、解答候補として出力する。

【０２０８】

以下に、固有表現抽出の一般的な手法の例について説明する。

（１）機械学習を用いる手法

機械学習を用いて固有表現を抽出する手法がある（例えば、以下の文献（１２）参照）。

【０２０９】

文献（１２）：浅原正幸，松本裕治，日本語固有表現抽出における冗長的な形態素解析の利用情報処理学会自然言語処理研究会 NL153-7 2002

まず、例えば、「日本の首相は小泉さんです。」という文を、各文字に分割し、分割した文字について、以下のように、 Ｂ－LOCATION、 Ｉ－LOCATION等の正解タグを付与することによって、正解を設定する。以下の一列目は、分割された各文字であり、各文字の正解タグは二列目である。

日 Ｂ－LOCATION
本 Ｉ－LOCATION
の 0
首 0
相 0
は 0
小 Ｂ－PERSON
泉 Ｉ－PERSON
さ 0
ん 0
で 0

す O
。 O
上記において、B －？？？は、ハイフン以下の固有表現の種類の始まりを意味するタグである。例えば、 B－LOCATIONは、地名という固有表現の始まりを意味しており、 B－PERSONは、人名という固有表現の始まりを意味している。また、I －？？？は、ハイフン以下の固有表現の種類の始まり以外を意味するタグであり、O はこれら以外である。従って、例えば、文字「日」は、地名という固有表現の始まりに該当する文字であり、文字「本」までが地名という固有表現である。

【０２１０】

このように、各文字の正解を設定しておき、このようなデータから学習し、新しいデータでこの正解を推定し、この正解のタグから、各固有表現の始まりと、どこまでがその固有表現かを認識して、固有表現を推定する。

【０２１１】

この各文字に設定された正解のデータから学習するときには、システムによってさまざまな情報を素性という形で利用する。例えば、

日 B－LOCATION

の部分は、

日本－B 名詞－B

などの情報を用いる。日本－B は、日本という単語の先頭を意味し、名詞－B は、名詞の先頭を意味する。単語や品詞の認定には、例えば前述したChaSenによる形態素解析を用いる。ChaSenを用いれば、入力された日本語を単語に分割することができる。例えば、ChaSenは、前述したように、日本語文を分割し、さらに、各単語の品詞も推定してくれる。例えば、「学校へ行く」を入力すると以下の結果を得ることができる。

【０２１２】

学校　ガッコウ　学校　名詞－一般
へ　へ　へ　助詞－格助詞－一般
行く　イク　行く　動詞－自立　五段・カ行促音便　基本形
ＥＯＳ

このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。

【０２１３】

なお、例えば、上記の文献（１２）では、素性として、入力文を構成する文字の、文字自体（例えば、「小」という文字）、字種（例えば、ひらがなやカタカナ等）、品詞情報、タグ情報（例えば、「 B－PERSON」等）を利用している。

【０２１４】

これら素性を利用して学習する。タグを推定する文字やその周辺の文字にどういう素性が出現するかを調べ、どういう素性が出現しているときにどういうタグになりやすいかを学習し、その学習結果を利用して新しいデータでのタグの推定を行なう。機械学習には、例えばサポートベクトルマシンを用いる。

【０２１５】

固有表現抽出には、上記の手法の他にも種々の手法がある。例えば、最大エントロピーモデルと書き換え規則を用いて固有表現を抽出する手法がある（文献（１３）参照）。

【０２１６】

文献（１３）：内元清貴，馬青，村田真樹，小作浩美，内山将夫，井佐原均，最大エントロピーモデルと書き換え規則に基づく固有表現抽出，言語処理学会誌，Vol.7，No.2，2000 参照）。

【０２１７】

また、例えば、以下の文献（１４）に、サポートベクトルマシンを用いて日本語固有表現抽出を行う手法について記載されている。

【０２１８】

文献（１４）：山田寛康，工藤拓，松本裕治，Support Vector Machineを用いた日本語固有表現抽出，情報処理学会論文誌，Vol.43，No.1"，2002

（２）形態素解析を用いる手法

形態素解析システム（例えば、前述したChaSen）を用いれば、入力された日本語を単語に分割することができる。

【０２１９】

例えば、ChaSenは、前述したように、日本語文を分割し、さらに、各単語の品詞も推定してくれる。例えば、「学校へ行く」を入力すると以下の結果を得ることができる。

【０２２０】

学校　ガッコウ　学校　名詞－一般
へ　へ　へ　助詞－格助詞－一般
行く　イク　行く　動詞－自立　五段・カ行促音便　基本形
ＥＯＳ

このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。

【０２２１】

具体的には、
入力：
日本の首都は東京です
出力：
日本ニッポン日本名詞－固有名詞－地域－国
のノの助詞－連体化
首都シュト首都名詞－一般
はハは助詞－係助詞
東京トウキョウ東京名詞－固有名詞－地域－一般
ですデスです助動詞特殊・デス基本形
EOS
は chasen の出力であり、名詞－固有名詞－地域という品詞が出力される。
このシステムを使って、例えば地名の固有表現を取り出すことができる。

【０２２２】

また、
入力：
村山首相が言った
出力：
村山ムラヤマ村山名詞－固有名詞－人名－姓
首相シュショウ首相名詞－一般
がガが助詞－格助詞－一般
言っイッ言う動詞－自立五段・ワ行促音便連用タ接続
たタた助動詞特殊・タ基本形
EOS
も chasen の出力であるが、これだと名詞－固有名詞－人名という品詞が出力される。このシステムを使って、例えば人名の固有表現を取り出すことができる。

（３）作成したルールを用いる手法

人手でルールを作って固有表現を取り出すという方法もある。

【０２２３】

例えば、
名詞＋「さん」だと人名とする
名詞＋「首相」だと人名とする
名詞＋「町」だと地名とする
名詞＋「市」だと地名とする

などである。

【０２２４】

図５は、本発明の第１の実施の形態における質問応答処理フローの一例を示す図である。キーワード入力部１１に、第１のキーワードと第２のキーワードを入力キーワードとして入力する（ステップＳ１）。例えば、第１のキーワード「日本」と、第２のキーワード「面積」とを入力する。

【０２２５】

キーワード増加部１２のパターン抽出部１２１で、入力キーワードをキーワード抽出用ＤＢ１６で全文検索し、入力キーワードの周辺に出現したパターンをc_iとして抽出する（ステップＳ２）。周辺に出現するパターンの定義は適宜行なう。パターンc_iの抽出は、第１のキーワードと第２のキーワードそれぞれについて行う。

【０２２６】

キーワード増加部１２のキーワード抽出部１２２で、パターン抽出部１２１で抽出したパターンc_iをキーワード抽出用ＤＢ１６で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートし、キーワードとして出力する（ステップＳ３）。ステップＳ３の処理によって、例えば、第１のキーワードが、「日本」、「アメリカ」、「ドイツ」という３つの第３のキーワードに増加し、第２のキーワードが、「面積」、「人口」、「緯度」の３つの第４のキーワードに増加する。

【０２２７】

次に、質問作成部１３が、出力されたキーワードにより構成される質問を作成する（ステップＳ４）。ステップＳ４においては、第３のキーワードと第４のキーワードとにより構成される質問を作成する。例えば、質問作成部１３は、第３のキーワード「アメリカ」と第４のキーワード「人口」とにより構成される質問「アメリカの人口は？」を作成する。質問応答装置１が質問作成部１３を備えない構成を採るときは、上記ステップＳ４の処理は、省略される。

【０２２８】

次に、解答候補抽出部１４は、作成された各質問に対する解答候補を、上述した機械学習の手法を用いて抽出する（ステップＳ５）。質問応答装置１が質問作成部１３を備えない構成を採るときは、上記ステップＳ５において、解答候補抽出部１４は、キーワード増加部１２によって出力されたキーワードによって構成される質問に対する解答候補を、機械学習の手法を用いて抽出する。そして、解答表出力部１５が、解答表を出力する（ステップＳ６）。

【０２２９】

図６は、本発明の第１の実施の形態における質問応答装置の構成の別の例を示す図である。質問応答装置１０は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図６中に示す質問応答装置１０が備える構成要素のうち、図１に示す質問応答装置１が備える構成要素と同一の符号が付けられたものは、当該質問応答装置１が備える構成要素と同様の機能を有する。

【０２３０】

本発明の実施の形態においては、図６に示す構成から質問作成部１３を省略し、解答候補抽出部１４が、キーワード増加部６０によって出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０２３１】

質問応答装置１０のキーワード増加部６０は、キーワード入力部１１に入力されたキーワードを増加させる。すなわち、キーワード増加部６０は、例えば、キーワード入力部１１に入力された第１のキーワードに基づいて、第１のキーワードの数より多い第３のキーワードを出力する。また、キーワード増加部６０は、例えば、キーワード入力部１１に入力された第２のキーワードに基づいて、第２のキーワードの数より多い第４のキーワード

を出力する。

【０２３２】

単語データデータベース（ＤＢ）６１には、単語と単語の分野との対応情報が格納されている。例えば、図７に示すような、単語と単語の分野との対応情報が格納されている。例えば、「国名」という分野に対応する単語として、日本、アメリカ、ドイツ、・・・といった単語が格納されている。

【０２３３】

また、シソーラスデータベース（ＤＢ）６２には、意味的類似による単語の分類情報であるシソーラスデータが格納されている。例えば、シソーラスＤＢ６２には、図８に示すような、単語と単語に振られた１０桁の数字（分類番号）との対応情報がシソーラスデータとして格納されている。図８に示す例では、シソーラスデータが分類語彙表の形式で示されている。

【０２３４】

なお、分類語彙表とは、一般に、単語を意味に基づいて整理した表であり、各単語に対して分類番号という数字が付与されている。この１０桁の分類番号は、７レベルの階層構造を示しており、上位５レベルは分類番号の最初の５桁で表現され、６レベル目は次の２桁、最下層のレベルは最後の３桁で表現されている。

【０２３５】

類似度算出部１００は、シソーラスＤＢ６２中のシソーラスデータに基づいて、キーワード入力部１１に入力されたキーワードとシソーラスデータ中の単語との類似度を算出する。キーワード抽出部１０１は、例えば、算出された類似度が予め定めた閾値以上の単語をキーワードとして抽出し、出力する。また、キーワード抽出部１０１は、例えば、算出された類似度が大きい順に所定の個数の単語をシソーラスデータ中から取り出して、キーワードとして出力する構成を採ることもできる。

【０２３６】

本発明の実施の形態においては、キーワード抽出部１０１は、単語データＤＢ６１中に格納された、単語と単語の分野との対応情報に基づいて、キーワード入力部１１に入力されたキーワードと同じ分野の単語をキーワードとして抽出し、出力する構成を採ることもできる。

【０２３７】

上記の質問応答装置１０を用いた場合の質問応答処理フローは、図５に示す質問応答処理フローと、ステップＳ２、ステップＳ３の処理が異なる以外は、同様である。質問応答装置１０を用いた場合の質問応答処理フローの一例においては、図５のステップＳ２およびステップＳ３の代わりに、キーワード増加部６０のキーワード抽出部１０１で、キーワード入力部１１に入力されたキーワードと同じ分野の単語を単語データＤＢ６１中から抽出し、キーワードとして出力する。

【０２３８】

例えば、キーワード入力部１１に第１のキーワード「日本」が入力されたとすると、キーワード抽出部１０１は、図７に示す単語データＤＢ６１から、単語「日本」が対応する「国名」という分野に属する（対応する）単語である「日本」、「アメリカ」、「ドイツ」、・・・を抽出し、第３のキーワードとして出力する。また、例えば、キーワード入力部１１に第２のキーワード「面積」が入力されたとすると、キーワード抽出部１０１は、図７に示す単語データＤＢ６１から、単語「面積」が対応する「数値表現」という分野に属する（対応する）単語である「面積」、「人口」、「緯度」、・・・を抽出し、第４のキーワードとして出力する。

【０２３９】

また、質問応答装置１０を用いた場合の質問応答処理フローの別の例においては、図５のステップＳ２およびステップＳ３の代わりに、例えば、キーワード増加部６０の類似度算出部１００が、キーワード入力部１１に入力されたキーワードとシソーラスＤＢ６２中の単語との類似度を算出し、キーワード増加部６０のキーワード抽出部１０１が、算出さ

れた類似度が予め定めた閾値以上の単語をキーワードとして出力する。

【０２４０】

なお、例えば、キーワード抽出部１０１は、算出された類似度が大きい順に所定の個数の単語をシソーラスデータ中から取り出して、キーワードとして出力する構成を採ることもできる。

【０２４１】

類似度算出部１００は、入力されたキーワードとシソーラスＤＢ６２中の単語との類似度を、例えば以下のようにして算出する。図８に示すシソーラスＤＢ６２内に格納されたシソーラスデータ（分類語彙表）中の各単語に振られた、１０桁の分類番号における各桁の数字の一致の割合を用いて、類似度を求める。すなわち、例えば、分類語彙表中の各単語に振られた分類番号について、キーワード入力部１１に入力されたキーワードと同一の単語に振られた分類番号との間での、各桁の数字の一致の割合を算出し、算出された値を類似度とする。なお、例えば、分類番号の６桁目と７桁目、および、８桁目と９桁目と１０桁目は、それぞれ連続した１つの数字として考える。

【０２４２】

例えば、キーワード入力部１１に第１のキーワードとして入力されたキーワードが「日本」である場合、図８に示す分類語彙表中の単語「日本」と「アメリカ」には、それぞれ以下のような分類番号が振られている。以下では、分類番号の上位５レベルと、６レベル目と、最下層のレベルとの間を空白で区切って示す。

【０２４３】

日本　　：１２５９０　０１　０１２

アメリカ：１２５９０　０４　１９２

例えば、両単語の分類番号の上位５レベルにおいて、最初の５桁が一致するので、算出されるキーワード「日本」と分類語彙表中の単語「アメリカ」との類似度は、類似度５である。

【０２４４】

また、例えば、キーワード入力部１１に第２のキーワードとして入力されたキーワードが「面積」である場合、分類語彙表中の単語「面積」と「人口」には、それぞれ以下のような分類番号が振られている。

【０２４５】

面積：１２６３０　１３　０１５

人口：１２６３０　１０　０１２

例えば、両単語の分類番号の上位５レベルにおいて、最初の５桁が一致するので、算出されるキーワード「面積」と分類語彙表中の単語「人口」との類似度は、類似度５である。

【０２４６】

また、例えば、キーワード入力部１１に第２のキーワードとして入力されたキーワードが「人口」である場合、分類語彙表中の単語「人口」と「緯度」には、それぞれ以下のような分類番号が振られている。

【０２４７】

人口：１２６３０　１０　０１２

緯度：１２６３０　１０　０１５

例えば、両単語の分類番号の上位５レベルにおいて、最初の５桁が一致し、また、６レベル目の２桁の数字「１０」が一致するので、算出されるキーワード「人口」と分類語彙表中の単語「緯度」との類似度は、類似度７である。

【０２４８】

また、例えば、キーワード入力部１１に第２のキーワードとして入力されたキーワードが「人口」である場合、分類語彙表中の単語「人口」と「アメリカ」には、それぞれ以下のような分類番号が振られている。

【０２４９】

人口　　：１２６３０　１０　０１２
アメリカ：１２５９０　０４　１９２

例えば、両単語の分類番号の上位５レベルにおいて、最初の２桁が一致するため、算出されるキーワード「人口」と分類語彙表中の単語「アメリカ」との類似度は、類似度２である。

【０２５０】

図９は、本発明の第２の実施の形態における質問応答装置の構成の一例を示す図である。第２の実施の形態においては、例えば、第１のキーワード「日本」と、第２のキーワード「首都」＋疑問代名詞「はどこですか？」が入力されると、第１のキーワード「日本」に基づいて、第１のキーワードを、例えば「日本」、「アメリカ」、「ドイツ」の３つに増加させる。そして、増加後の第１のキーワードと、第２のキーワード「首都」＋疑問代名詞「はどこですか？」により構成される、例えば「日本の首都はどこですか？」、「アメリカの首都はどこですか？」、「ドイツの首都はどこですか？」という各質問に対する解答を出力する。

【０２５１】

質問応答装置２は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図９に示す質問応答装置２の構成要素のうち、キーワード入力部１１、解答表出力部１５、キーワード抽出用ＤＢ１６、パターン抽出部１２１、キーワード抽出部１２２は、それぞれ、図１に示す質問応答装置１の、同符号の構成要素と同様である。本発明の実施の形態においては、図９に示す構成から後述する質問作成部２３を省略し、解答候補抽出部２４が、キーワード増加部１８によって出力されたキーワードによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０２５２】

キーワード入力部１１には、キーワードが入力される。例えば、第１のキーワード「日本」と第２のキーワード「首都」が入力される。疑問代名詞入力部２１には、キーワード入力部１１に入力された第２のキーワードに対応付けられた疑問代名詞が入力される。例えば、「はどこですか？」という疑問代名詞が入力される。この他、疑問代名詞入力部２１に入力される疑問代名詞として、例えば、「は何時ですか？」、「は誰ですか？」などが挙げられる。なお、疑問代名詞入力部２１に入力される疑問代名詞は、ユーザの指定入力に基づいて入力されるものであってもよいし、また、質問応答装置２とは別のコンピュータによって入力されるものであってもよい。

【０２５３】

解答タイプ推定部２２は、疑問代名詞入力部２１に入力された疑問代名詞に基づいて、後述する質問作成部２３によって作成される質問、または、質問応答装置２が質問作成部２３を備えない構成を採るときは、後述するキーワード増加部１８によって出力されたキーワードによって構成される質問に対する解答候補の言語表現の類型である解答タイプを推定する。例えば、入力された疑問代名詞が「はどこですか？」である場合には、解答タイプは「固有名詞（地名）」であると推定する。本発明の実施の形態においては、解答タイプ推定部２２は、疑問代名詞入力部２１に入力された疑問代名詞ではなく、予め定められた疑問代名詞に基づいて、上記解答タイプを推定してもよい。

【０２５４】

キーワード増加部１８は、キーワード抽出技術を用いて、入力された第１のキーワードと同じ分野のキーワードをキーワード抽出用ＤＢ１６から抽出して、第１のキーワードを増加させ、第３のキーワードとして出力する。第２の実施の形態では、キーワード増加部１８は、第２のキーワード（例えば、「首都」）については増加させずに、質問作成部２３に対して出力する。質問応答装置２が質問作成部２３を備えない構成を採るときは、キーワード増加部１８は、第３のキーワードと第２のキーワードを解答候補抽出部２４に対して出力する。

【０２５５】

質問作成部２３は、キーワード増加部１８の処理によって出力された第３のキーワードと、第２のキーワードと、疑問代名詞入力部２１に入力された疑問代名詞（または予め定められた疑問代名詞）とに基づいて、複数の質問を作成する。

【０２５６】

知識データベース（ＤＢ）２５には、解答候補の検索対象となる文書データ群が蓄積される。蓄積される文書データ群としては、例えば、新聞記事データ・百科事典データなどの文書データ群が挙げられる。

【０２５７】

解答候補抽出部２４は、知識ＤＢ２５から、質問作成部２３によって作成された各質問を構成するキーワード（または、キーワード増加部１８から出力された第３のキーワードと第２のキーワード）を含む文書データを検索し、この検索処理で抽出された文書データから、解答タイプ推定部２２によって推定された解答タイプに適合する言語表現を、解答候補として抽出する。

【０２５８】

解答表出力部１５は、抽出された各解答候補が質問と対応付けられた表を解答表として出力する。例えば、図１０に示すような解答表を出力する。

【０２５９】

図１０に示す解答表においては、例えば、「日本の首都はどこですか？」という質問に対する解答として、データ項目「日本」に対応する行とデータ項目「首都」に対応する列とが交差する枡目に、「東京」が格納され、「アメリカの首都はどこですか？」という質問に対する解答として、データ項目「アメリカ」に対応する行とデータ項目「首都」に対応する列とが交差する枡目に、「ワシントン」が格納され、「ドイツの首都はどこですか？」という質問に対する解答として、データ項目「アメリカ」に対応する行とデータ項目「首都」に対応する列とが交差する枡目に、「ベルリン」が格納される。

【０２６０】

図１１は、本発明の第２の実施の形態における質問応答処理フローの一例を示す図である。キーワード入力部１１に、第１のキーワードと第２のキーワードを入力キーワードとして入力する（ステップＳ１１）。例えば、第１のキーワード「日本」と第２のキーワード「首都」が入力される。また、疑問代名詞入力部２１に、第２のキーワードに対応付けられた疑問代名詞が入力される（ステップＳ１２）。例えば、第２のキーワード「首都」に対応付けられた疑問代名詞「はどこですか？」が入力される。

【０２６１】

キーワード増加部１８のパターン抽出部１２１で、第１のキーワードをキーワード抽出用ＤＢ１６で全文検索し、第１のキーワードの周辺に出現したパターンをc_iとして抽出する（ステップＳ１３）。周辺に出現するパターンの定義は適宜行なう。

【０２６２】

キーワード増加部１８のキーワード抽出部１２２で、パターン抽出部１２１で抽出したパターンc_iをキーワード抽出用ＤＢ１６で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートし、第３のキーワードとして出力する（ステップＳ１４）。ステップＳ１４の処理によって、例えば、第１のキーワードが、「日本」、「アメリカ」、「ドイツ」という３つの第３のキーワードに増加する。

【０２６３】

解答タイプ推定部２２が、疑問代名詞入力部２１に入力された疑問代名詞に基づいて、解答タイプを推定する（ステップＳ１５）。例えば、入力された疑問代名詞が「はどこですか？」である場合には、解答タイプ推定部２２は、解答タイプが「固有名詞（地名）」であると推定する。

【０２６４】

質問作成部２３が、疑問代名詞入力部２１に入力された疑問代名詞を用いて、第３のキーワードと第２のキーワードとにより構成される質問を作成する（ステップＳ１６）。例

えば、質問作成部２３は、第３のキーワード「アメリカ」と第２のキーワード「首都」とにより構成される質問「アメリカの首都はどこですか？」を作成する。質問応答装置２が質問作成部２３を備えない構成を採るときは、上記ステップＳ１６の処理は、省略される。

【０２６５】

次に、解答候補抽出部２４は、作成された各質問に対する解答候補を抽出する（ステップＳ１７）。すなわち、解答候補抽出部２４は、知識ＤＢ２５から、質問作成部２３によって作成された各質問を構成するキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、解答タイプ推定部２２によって推定された解答タイプに適合する言語表現を、解答候補として抽出する。質問応答装置２が質問作成部２３を備えない構成を採るときは、上記ステップＳ１７において、解答候補抽出部２４は、知識ＤＢ２５から、キーワード増加部１８によって出力されたキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、解答タイプ推定部２２によって推定された解答タイプに適合する言語表現を、キーワード増加部１８によって出力されたキーワードによって構成される質問に対する解答候補として抽出する。そして、解答表出力部１５が、解答表を出力する（ステップＳ１８）。例えば、上述した図１０に示すような解答表が出力される。

【０２６６】

図１２は、本発明の第２の実施の形態の変形例１の構成例を示す図である。質問応答装置２０は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図１２中に示す質問応答装置１０が備える構成要素のうち、図６に示す質問応答装置１０が備える構成要素または図９に示す質問応答装置２が備える構成要素と同一の符号が付けられたものは、当該質問応答装置１０または質問応答装置２が備える構成要素と同様の機能を有する。

【０２６７】

質問応答装置２０のキーワード増加部６３は、キーワード入力部１１に入力された第１のキーワードを増加させて、第３のキーワードとして出力する。また、キーワード入力部１１に入力された第２のキーワードについては、増加させずに、質問作成部２３に対して出力する。本発明の実施の形態においては、図１２に示す構成から質問作成部２３を省略し、解答候補抽出部２４が、キーワード増加部６３によって出力された第３のキーワードと第２のキーワードとによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０２６８】

上記の質問応答装置２０を用いた場合の質問応答処理フローは、図１１に示す質問応答処理フローと、ステップＳ１３、ステップＳ１４の処理が異なる以外は、同様である。質問応答装置１０を用いた場合の質問応答処理フローの一例においては、図１１のステップＳ１３およびステップＳ１４の代わりに、キーワード増加部６３のキーワード抽出部１０１で、キーワード入力部１１に入力された第１のキーワードと同じ分野の単語を単語データＤＢ６１中から抽出し、第３のキーワードとして出力する。

【０２６９】

また、質問応答装置１０を用いた場合の質問応答処理フローの別の例においては、図１１のステップＳ１３およびステップＳ１４の代わりに、キーワード増加部６３の類似度算出部１００が、キーワード入力部１１に入力された第１のキーワードとシソーラスＤＢ６２中の単語との類似度を算出し、キーワード増加部６３のキーワード抽出部１０１が、算出された類似度が予め定めた閾値以上の単語を第３のキーワードとして出力する。

【０２７０】

また、キーワード抽出部１０１は、例えば、算出された類似度が大きい順に所定の個数の単語をシソーラスデータ中から取り出して、第３のキーワードとして出力する構成を採ることもできる。

【０２７１】

本発明の第２の実施の形態の変形例２においては、図９に示す質問応答装置２または図１２に示す質問応答装置２０において、疑問代名詞入力部２１には、キーワード入力部１１に入力されるキーワードと対応付けられていない疑問代名詞が入力される。質問応答装置２のキーワード増加部１８（または質問応答装置２０のキーワード増加部６３）は、キーワード入力部１１に入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、キーワード入力部１１に入力された第２のキーワードに基づいて、第４のキーワードを出力キーワードとして出力する。解答タイプ推定部２２は、疑問代名詞入力部２１に入力された疑問代名詞に基づいて、解答タイプを推定する。そして、解答候補抽出部２４は、知識ＤＢ２５から、キーワード増加部１８によって出力されたキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、解答タイプ推定部２２によって推定された解答タイプに適合する言語表現を、キーワード増加部１８によって出力されたキーワードによって構成される質問に対する解答候補として抽出する。そして、解答表出力部１５が、解答表を出力する。なお、上記の本発明の第２の実施の形態の変形例２においては、解答タイプ推定部２２は、疑問代名詞入力部２１に入力された疑問代名詞ではなく、予め定められた疑問代名詞に基づいて解答タイプを推定する構成を採ってもよい。

【０２７２】

図１３は、本発明の第３の実施の形態における質問応答装置の構成の一例を示す図である。第３の実施の形態では、第１の実施の形態のような機械学習の手法を用いるのではなく、入力された解答タイプ（または予め定められた解答タイプ）を用いて解答候補を抽出する。

【０２７３】

第３の実施の形態においては、例えば、第１のキーワード「日本」と第２のキーワード「首都」と、解答タイプ「固有名詞（地名）」が入力されると、第１のキーワード「日本」に基づいて、第１のキーワードを、例えば「日本」、「アメリカ」、「ドイツ」という３つの第３のキーワードに増加させる。また、第２のキーワード「首都」に基づいて、第２のキーワードを、例えば「首都」、「旧首都」、「最南端都市」という３つの第４のキーワードに増加させる。

【０２７４】

そして、増加後の第３のキーワードと第４のキーワードとの組み合わせにより構成される、例えば「日本の首都は？」、「アメリカの旧首都は？」、「ドイツの最南端都市は？」・・・といった各質問に対する解答を出力する。より具体的には、後述するように、「日本の首都は？」という質問を構成する第３のキーワード「日本」と第４のキーワード「首都」を、解答候補の検索対象となる文書データ群から検索し、両キーワードを含む文書中の言語表現を解答候補として抽出するとともに、抽出された解答候補のうち解答タイプ「固有名詞（地名）」に適合するものを解答として出力する。

【０２７５】

質問応答装置３は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図１３に示す質問応答装置３の構成要素のうち、キーワード入力部１１、キーワード増加部１２、質問作成部１３、解答表出力部１５、キーワード抽出用ＤＢ１６、パターン抽出部１２１、キーワード抽出部１２２は、それぞれ、図１に示す質問応答装置１の、同符号の構成要素と同様であり、解答候補抽出部２４、知識ＤＢ２５は、図９に示す質問応答装置２の、同符号の構成要素と同様である。本発明の実施の形態においては、図１３に示す構成から質問作成部１３を省略し、解答候補抽出部２４が、キーワード増加部１２によって出力されたキーワードによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０２７６】

キーワード入力部１１には、キーワードが入力される。例えば、第１のキーワード「日本」と第２のキーワード「首都」が入力される。解答タイプ入力部３１には、質問作成部１３によって作成される質問、または、質問応答装置３が質問作成部１３を省略する構成

を採るときは、キーワード増加部１２によって出力されるキーワードによって構成される質問に対する解答候補の解答タイプが入力される。例えば、「固有名詞（地名）」という解答タイプが入力される。

【０２７７】

この他、解答タイプ入力部３１に入力される解答タイプとして、例えば、「固有名詞（数値）」、「固有名詞（人名）」、「カタカナ表現」（カタカナだけで表現されるもの）、「名詞」、「動詞」などが挙げられる。なお、解答タイプ入力部３１に入力される解答タイプは、ユーザの指定入力に基づいて入力されるものであってもよいし、また、質問応答装置３とは別のコンピュータによって入力されるものであってもよい。

【０２７８】

キーワード増加部１２は、図１を参照して説明したように、キーワード抽出技術を用いて、入力された各キーワードと同じ分野のキーワードをキーワード抽出用ＤＢ１６から抽出して、キーワードを増加させる。

【０２７９】

例えば、キーワード増加部１２は、入力された第１のキーワードに基づいて、第１のキーワードの数より多い第３のキーワードを出力する。また、例えば、キーワード増加部１２は、入力された第２のキーワードに基づいて、第２のキーワードの数より多い第４のキーワードを出力する。

【０２８０】

質問作成部１３は、第３のキーワードと第４のキーワードとによって構成される質問を複数作成する。解答候補抽出部２４は、知識ＤＢ２５から、質問作成部１３によって作成された各質問を構成するキーワード（または、キーワード増加部１２によって出力された第３のキーワードと第４のキーワード）を含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、解答タイプ入力部３１に入力された解答タイプ（または予め定められた解答タイプ）に適合する言語表現を、解答候補として抽出する。

【０２８１】

解答表出力部１５は、抽出された各解答候補が質問と対応付けられた表を解答表として出力する。

【０２８２】

図１４は、本発明の第３の実施の形態における質問応答処理フローの一例を示す図である。キーワード入力部１１に、第１のキーワードと第２のキーワードを入力キーワードとして入力する（ステップＳ２１）。例えば、第１のキーワード「日本」と第２のキーワード「首都」が入力される。また、解答タイプ入力部３１に、質問作成部１３により作成される質問に対する解答候補の解答タイプを入力する（ステップＳ２２）。例えば、解答タイプとして、「固有名詞（地名）」が入力される。なお、質問応答装置３が質問作成部１３を備えない構成を採るときは、解答タイプ入力部３１には、キーワード増加部１２によって出力されるキーワードによって構成される質問に対する解答候補の解答タイプが入力される。

【０２８３】

キーワード増加部１２のパターン抽出部１２１で、入力キーワードをキーワード抽出用ＤＢ１６で全文検索し、複数の入力キーワードの周辺に出現したパターンをc_iとして抽出する（ステップＳ２３）。周辺に出現するパターンの定義は適宜行なう。パターンc_iの抽出は、第１のキーワードと第２のキーワードそれぞれについて行う。

【０２８４】

キーワード増加部１２のキーワード抽出部１２２で、パターン抽出部１２１で抽出したパターンc_iをキーワード抽出用ＤＢ１６で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートし、キーワードとして出力する（ステップＳ２４）。ステップＳ２４の処理によって、例えば、第１のキーワードが、「日本」、「アメリカ」、「ドイツ」という３つの第３のキーワードに増加する。また、第２のキーワードが、「首都」、「旧首都」、「最南端

都市」という３つの第４のキーワードに増加する。

【０２８５】

質問作成部２３が、出力されたキーワードにより構成される質問を作成する（ステップＳ２５）。ステップＳ２５においては、出力された第３のキーワードと第４のキーワードとにより構成される質問を作成する。例えば、質問作成部２３は、第３のキーワード「アメリカ」と第４のキーワード「首都」とにより構成される質問「アメリカの首都は？」を作成する。質問応答装置３が質問作成部１３を備えない構成を採るときは、上記ステップＳ２５の処理は、省略される。

【０２８６】

次に、解答候補抽出部２４は、作成された各質問に対する解答候補を抽出する（ステップＳ２６）。すなわち、解答候補抽出部２４は、知識ＤＢ２５から、質問作成部１３によって作成された各質問を構成するキーワードを含む文書データを検索し、この検索処理で抽出された文書データから、解答タイプ入力部３１に入力された解答タイプに適合する言語表現を、解答候補として抽出する。質問応答装置３が質問作成部１３を備えない構成を採るときは、上記ステップＳ２６において、解答候補抽出部２４は、知識ＤＢ２５から、キーワード増加部１２によって出力されたキーワード（第３のキーワードと第４のキーワード）を含む文書データを検索し、この検索処理で抽出された文書データから、解答タイプ入力部３１に入力された解答タイプに適合する言語表現を、キーワード増加部１２によって出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補として抽出する。そして、解答表出力部１５が、解答表を出力する（ステップＳ２７）。例えば、図１５に示すような解答表が出力される。

【０２８７】

図１６は、本発明の第３の実施の形態の変形例１の構成例を示す図である。質問応答装置３０は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図１６中に示す質問応答装置３０が備える構成要素のうち、図１に示す質問応答装置１が備える構成要素または図６に示す質問応答装置１０または図１３に示す質問応答装置３が備える構成要素と同一の符号が付けられたものは、当該質問応答装置１または質問応答装置１０または質問応答装置３が備える構成要素と同様の機能を有する。

【０２８８】

本発明の実施の形態においては、図１６に示す構成から質問作成部１３を省略し、解答候補抽出部２４が、キーワード増加部６０によって出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０２８９】

上記の質問応答装置３０を用いた場合の質問応答処理フローは、図１４に示す質問応答処理フローと、ステップＳ２３、ステップＳ２４の処理が異なる以外は、同様である。質問応答装置３０を用いた場合の質問応答処理フローの一例においては、図１４のステップＳ２３およびステップＳ２４の代わりに、キーワード増加部６０のキーワード抽出部１０１で、キーワード入力部１１に入力された第１のキーワードと同じ分野の単語を単語データＤＢ６１中から抽出し、第３のキーワードとして出力する。また、キーワード抽出部１０１で、キーワード入力部１１に入力された第２のキーワードと同じ分野の単語を単語データＤＢ６１中から抽出し、第４のキーワードとして出力する。

【０２９０】

また、質問応答装置３０を用いた場合の質問応答処理フローの別の例においては、図１４のステップＳ２３およびステップＳ２４の代わりに、キーワード増加部６０の類似度算出部１００が、キーワード入力部１１に入力された第１のキーワードとシソーラスＤＢ６２中の単語との類似度を算出し、キーワード増加部６０のキーワード抽出部１０１が、算出された類似度が予め定めた閾値以上の単語を第３のキーワードとして出力する。また、類似度算出部１００が、キーワード入力部１１に入力された第２のキーワードとシソーラ

スＤＢ６２中の単語との類似度を算出し、キーワード抽出部１０１が、算出された類似度が予め定めた閾値以上の単語を第４のキーワードとして出力する。

【０２９１】

また、キーワード抽出部１０１は、例えば、算出された類似度が大きい順に所定の個数の単語をシソーラスデータ中から取り出して、上記の第３のキーワード、第４のキーワードとして出力する構成を採ることもできる。

【０２９２】

本発明の第３の実施の形態の変形例２では、例えば、図１３に示す質問応答装置３または図１６に示す質問応答装置３０において、解答タイプ入力部３１には、キーワード増加部１２（またはキーワード増加部６０）によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプであって、キーワード入力部１１に入力された第２のキーワードに対応付けられた解答タイプが入力される。キーワード増加部１２（またはキーワード増加部６０）は、キーワード入力部１１に入力された第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力し、キーワード入力部１１に入力された第２のキーワードを出力キーワードとして出力する。解答候補抽出部２４は、知識ＤＢ２５から、キーワード増加部１２（またはキーワード増加部６０）によって出力されたキーワード（第３のキーワードと第２のキーワード）を含む文書データを検索し、この検索処理で抽出された文書データから、当該第２のキーワードに対応する解答タイプ入力部３１に入力された解答タイプに適合する言語表現を、キーワード増加部１２によって出力された第３のキーワードと当該第２のキーワードとによって構成される質問に対する解答候補として抽出する。そして、解答表出力部１５が、解答表を出力する。

【０２９３】

なお、本発明の第３の実施の形態の変形例３では、解答候補抽出部２４は、知識ＤＢ２５から、キーワード増加部１２（またはキーワード増加部６０）によって出力されたキーワード（第３のキーワードと第２のキーワード）を含む文書データを検索し、この検索処理で抽出された文書データから、予め定められた、当該第２のキーワードに対応付けられた解答タイプに適合する言語表現を、キーワード増加部１２によって出力された第３のキーワードと当該第２のキーワードとによって構成される質問に対する解答候補として抽出する構成を採ってもよい。

【０２９４】

本発明の第３の実施の形態の変形例３では、例えば、図１３に示す質問応答装置３または図１６に示す質問応答装置３０において、キーワード入力部１１には、第１のキーワードと、複数のグループによってグループ化された第２のキーワードとが入力される。例えば、第１のキーワード「日本」と、人名のグループに属する第２のキーワード「首相」、「市長」と、地名のグループに属する第２のキーワード「首都」、「旧首都」が入力される。解答タイプ入力部３１には、キーワード入力部１１に入力される第２のキーワードが属する各グループに対応付けられた解答タイプが入力される。例えば、人名のグループに対応する解答タイプとして、解答タイプ「人名」が入力され、地名のグループに対応する解答タイプとして、解答タイプ「地名」が入力される。

【０２９５】

キーワード増加部１２（またはキーワード増加部６０）は、第１のキーワードに基づいて、第３のキーワードを出力キーワードとして出力する。また、第２のキーワードに基づいて、当該第２のキーワードが属するグループ毎に、第４のキーワードを出力キーワードとして出力する。例えば、第３のキーワードとして、「日本」、「ドイツ」、「アメリカ」が出力される。また、例えば、人名のグループに属する第４のキーワードとして、「首相」、「市長」、「ノーベル賞受賞者」が出力され、地名のグループに属する第４のキーワードとして、「首都」、「旧首都」、「最南端都市」が出力される。

【０２９６】

解答候補抽出部２４は、キーワード増加部１２（またはキーワード増加部６０）によっ

て出力された第３のキーワードと、人名のグループに属する第４のキーワードとによって構成される質問に対する解答候補を、解答タイプ入力部３１に入力された解答タイプ「人名」を用いて抽出する。例えば、「ドイツのノーベル賞受賞者は？」という質問に対する解答候補は、解答タイプ「人名」を用いて抽出される。また、解答候補抽出部２４は、キーワード増加部１２（またはキーワード増加部６０）によって出力された第３のキーワードと、地名のグループに属する第４のキーワードとによって構成される質問に対する解答候補を、解答タイプ入力部３１に入力された解答タイプ「地名」を用いて抽出する。例えば、「アメリカの首都は？」という質問に対する解答候補は、解答タイプ「地名」を用いて抽出される。そして、解答表出力部１５が解答表を出力する。

【０２９７】

図１７は、本発明の第４の実施の形態における質問応答装置の構成の一例を示す図である。第４の実施の形態では、キーワードの類似関係を用いて解答候補を抽出する。

【０２９８】

第４の実施の形態においては、例えば、第１のキーワード「日本」、「アメリカ」、・・・と第２のキーワード「面積」、「首都」、・・・と解答タイプ「固有名詞（数値）」、「固有名詞（地名）」、・・・が入力される。入力される解答タイプは、入力された第２のキーワードのそれぞれに対応付けられている。例えば、第２のキーワード「面積」に対応付けられた解答タイプは「固有名詞（数値）」であり、第２のキーワード「首都」に対応付けられた解答タイプは「固有名詞（地名）」である。

【０２９９】

第４の実施の形態では、例えば、入力された第１のキーワード「日本」、「アメリカ」、・・・に基づいて、第１のキーワードを多数の第３のキーワード（例えば「日本」、「アメリカ」、「ドイツ」、・・・）に増加させる。また、キーワード増加部１２が第２のキーワード「面積」、「首都」、・・・に基づいて、第２のキーワードを多数の第４のキーワード（例えば「面積」、「首都」、「旧首都」、・・・）に増加させる。

【０３００】

次に、第２のキーワード（と同一の第４のキーワード）のうち、第４のキーワードに類似するキーワードを、類似キーワードとして決定する。例えば、第４のキーワード「旧首都」に類似する第２のキーワード（と同一の第４のキーワード）「首都」を類似キーワードとして決定する。

【０３０１】

そして、第３のキーワードと第４のキーワードとの組み合わせにより構成される質問に対する解答の候補を、上記質問を構成する第４のキーワードに類似する類似キーワードに対応付けられている解答タイプを用いて抽出し、解答表を出力する。例えば、「日本の旧首都は？」という質問に対する解答の候補を、類似キーワード「首都」に対応付けられている解答タイプ「固有名詞（地名）」を用いて抽出し、解答表を出力する。

【０３０２】

質問応答装置４は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図１７に示す質問応答装置４の構成要素のうち、キーワード入力部１１、キーワード増加部１２、解答表出力部１５、キーワード抽出用ＤＢ１６、パターン抽出部１２１、キーワード抽出部１２２は、それぞれ、図１に示す質問応答装置１の、同符号の構成要素と同様であり、知識ＤＢ２５は、図９に示す質問応答装置２が備える知識ＤＢ２５と同様であり、解答タイプ入力部３１は、図１３に示す質問応答装置３が備える解答タイプ入力部３１と同様である。

【０３０３】

キーワード入力部１１には、キーワードが入力される。例えば、第１のキーワード「日本」、「アメリカ」、・・・と第２のキーワード「面積」、「首都」、・・が入力される。解答タイプ入力部３１には、質問作成部４２によって作成される質問、または、質問応答装置４が質問作成部４２を省略する構成を採るときは、キーワード増加部１２によって出力されるキーワードによって構成される質問に対する解答候補の解答タイプが入力され

る。入力される解答タイプは、特に、第２のキーワードに対応付けられている。

【０３０４】

例えば、解答タイプ入力部３１には、第２のキーワード「面積」に対応して、「固有名詞（数値）」という解答タイプが入力され、第２のキーワード「首都」に対応して、「固有名詞（地名）」という解答タイプが入力される。

【０３０５】

キーワード増加部１２は、図１を参照して説明したように、キーワード抽出技術を用いて、入力された各キーワードと同じ分野のキーワードをキーワード抽出用ＤＢ１６から抽出して、キーワードを増加させる。キーワード増加部１２の処理により、第１のキーワードから第３のキーワードが出力され、第２のキーワードから第４のキーワードが出力される。

【０３０６】

類似キーワード決定部４１は、各第４のキーワードに類似する、キーワード入力部１１に入力された第２のキーワード（と同一の第４のキーワード）を、類似キーワードとして決定する。類似キーワードの決定手法について以下に説明する。

【０３０７】

（共起ベクトルを用いる手法（１））

第４のキーワード毎に、キーワード増加部１２が抽出したパターンc_iと共起してキーワード抽出用ＤＢ１６中に出現した回数を算出し、算出した回数を要素とするベクトル（以下、「共起ベクトル」という）を求める。

【０３０８】

例えば、キーワード増加部１２におけるキーワード抽出処理において、第４のキーワード（１）がパターンc_1と共起して出現した回数が０、パターンc_2と共起して出現した回数が１、・・・、パターンc_nと共起して出現した回数が１とすると、第４のキーワード（１）についての共起ベクトルは、（０，１，・・・１）と求まる。同様にして、他の第４のキーワード（第２のキーワード（２）、第２のキーワード（３）、・・・）についての共起ベクトルを求める。

【０３０９】

キーワード入力部１１に入力された第２のキーワードと同一の第４のキーワードについての共起ベクトルと、対応する類似キーワードを求めたい第４のキーワードについての共起ベクトルとの類似の度合いを求める。例えば、キーワード入力部１１に入力された第２のキーワードと同一の第４のキーワードについての共起ベクトルが（a_1，a_2，a_3，・・・a_n）、対応する類似キーワードを求めたい第４のキーワードについての共起ベクトルが（b_1，b_2，b_3，・・・b_n）とすると、$(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2+\cdots+(a_n-b_n)^2$の値を算出する。算出された値が両共起ベクトル間の類似の度合いを示している。算出された値が低いほど、類似の度合いが高い。

【０３１０】

算出された値が最も低いときのキーワード入力部１１に入力された第２のキーワードと同一の第４のキーワードを、対応する類似キーワードを求めたい第４のキーワードに類似する類似キーワードとする。

【０３１１】

（共起ベクトルを用いる手法（２））

類似キーワード決定部４１は、まず、第４のキーワードを用いて知識ＤＢ２５を全文検索し、各第４のキーワードと共起して出現した語（共起語）を抽出する。そして、各第４のキーワードが、抽出された共起語と共起して知識ＤＢ２５中に出現した回数を要素とするベクトルを、各第４のキーワードについての共起ベクトルとして求める。

【０３１２】

例えば、第４のキーワード（１）が共起語w_1と共起して出現した回数が２、共起語w_2と共起して出現した回数が０、共起語w_3と共起して出現した回数が１・・・、パター

ンc_nと共起して出現した回数が１とすると、第４のキーワード（１）についての共起ベクトルは、（２，０，１，・・・１）と求まる。同様にして、他の第４のキーワード（第２のキーワード（２），第２のキーワード（３），・・・）についての共起ベクトルを求める。

【０３１３】

キーワード入力部１１に入力された第２のキーワードと同一の第４のキーワードについての共起ベクトルと、対応する類似キーワードを求めたい第４のキーワードについての共起ベクトルとの類似の度合いを求める。例えば、キーワード入力部１１に入力された第２のキーワードと同一の第４のキーワードについての共起ベクトルが（a_1，a_2，a_3，・・・a_n）、対応する類似キーワードを求めたい第４のキーワードについての共起ベクトルが（b_1，b_2，b_3，・・・b_n）とすると、$(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2+\cdots+(a_n-b_n)^2$の値を算出する。算出された値が両共起ベクトル間の類似の度合いを示している。算出された値が低いほど、類似の度合いが高い。

【０３１４】

算出された値が最も低いときのキーワード入力部１１に入力された第２のキーワードと同一の第４のキーワードを、対応する類似キーワードを求めたい第４のキーワードに類似する類似キーワードとする。

【０３１５】

なお、本発明の実施の形態においては、類似キーワード決定部４１が、知識ＤＢ２５ではなく、他の文書データを用いて上記共起ベクトルを求める構成を採ることもできる。例えば、大量の文書データが格納された大規模コーパス（図示を省略）を用いて上記共起ベクトルを求める構成を採ることもできる。

【０３１６】

（シソーラスデータを用いる手法）

シソーラスデータが分類語彙表の形式で格納されているシソーラスデータベース（図１７では図示を省略）を用意する。類似キーワード決定部４１は、シソーラスデータベース内に格納されているシソーラスデータ中の各単語に振られた、１０桁の分類番号における各桁の数字の一致の割合を用いて、第４のキーワードと、キーワード入力部１１に入力された第２のキーワード（と同一の第４のキーワード）との類似度を求める。

【０３１７】

すなわち、例えば、分類語彙表中の、対応する類似キーワードを求めたい第４のキーワードと同一の単語に振られた分類番号について、キーワード入力部１１に入力された第２のキーワード（と同一の第４のキーワード）と同一の単語に振られた分類番号との間での、各桁の数字の一致の割合を算出し、算出された値の大きさを類似度とする。そして、算出された値が最も大きいときの、上記第２のキーワード（と同一の第４のキーワード）を、対応する類似キーワードを求めたい第４のキーワードに類似する類似キーワードとして決定する。

【０３１８】

質問作成部４２は、キーワード増加部１２の処理によって出力された第３のキーワードと第４のキーワードとの組み合わせによって構成される質問を作成する。

【０３１９】

本発明の実施の形態においては、図１７に示す構成から質問作成部４２を省略し、解答候補抽出部４３が、キーワード増加部１２によって出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０３２０】

解答候補抽出部４３は、知識ＤＢ２５から、質問作成部４２によって作成された各質問を構成する第３のキーワードと第４のキーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、各質問を構成する第４のキーワードに類似する

類似キーワードに対応付けられて解答タイプ入力部３１に入力された解答タイプに適合する言語表現を、解答候補として抽出する。解答候補抽出部４３は、知識ＤＢ２５から、質問作成部４２によって作成された各質問を構成する第３のキーワードと第４のキーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、各質問を構成する第４のキーワードに類似する類似キーワードに予め対応付けられた解答タイプに適合する言語表現を、解答候補として抽出する構成を採ってもよい。

【０３２１】

また、解答候補抽出部４３は、質問応答装置４が質問作成部４２を省略する構成を採るときは、知識ＤＢ２５から、キーワード増加部１２によって出力される第３のキーワードと第４のキーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、第４のキーワードに類似する類似キーワードに対応付けられて解答タイプ入力部３１に入力された（または予め類似キーワードに対応付けられた）解答タイプに適合する言語表現を、第３のキーワードと当該第４のキーワードとによって構成される各質問に対する解答候補として抽出する。

【０３２２】

すなわち、各質問に対する解答候補の抽出に、各質問を構成する第４のキーワードに類似する類似キーワードに対応付けられた解答タイプを用いる。

【０３２３】

例えば、「日本の旧首都は？」という質問に対する解答の候補を、第４のキーワード「旧首都」に類似する類似キーワード「首都」に対応付けられている解答タイプ「固有名詞（地名）」を用いて抽出する。

【０３２４】

解答表出力部１５は、抽出された各解答候補が質問と対応付けられた表を解答表として出力する。

【０３２５】

図１８は、本発明の第４の実施の形態における質問応答処理フローの一例を示す図である。キーワード入力部１１に、第１のキーワードと第２のキーワードを入力キーワードとして入力する（ステップＳ３１）。例えば、第１のキーワード「日本」、「アメリカ」、・・・と第２のキーワード「面積」、「首都」、・・・が入力される。また、解答タイプ入力部３１に、第２のキーワードに対応付けられた解答タイプを入力する（ステップＳ３２）。例えば、第２のキーワード「面積」に対応付けられた解答タイプ「固有名詞（数値）」、第２のキーワード「首都」に対応付けられた解答タイプ「固有名詞（地名）」が入力される。

【０３２６】

キーワード増加部１２のパターン抽出部１２１で、入力キーワードをキーワード抽出用ＤＢ１６で全文検索し、複数の入力キーワードの周辺に出現したパターンをc_iとして抽出する（ステップＳ３３）。周辺に出現するパターンの定義は適宜行なう。なお、パターンc_iの抽出は、第１のキーワードと第２のキーワードそれぞれについて行う。

【０３２７】

キーワード増加部１２のキーワード抽出部１２２で、パターン抽出部１２１で抽出したパターンc_iをキーワード抽出用ＤＢ１６で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートし、キーワードとして出力する（ステップＳ３４）。

【０３２８】

ステップＳ３４の処理によって、例えば、第１のキーワードが、多数の第３のキーワード（例えば、「日本」、「アメリカ」、「ドイツ」、「イタリア」、「フランス」、「イギリス」・・・）に増加する。また、第２のキーワードが、多数の第４のキーワード（例えば、「面積」、「人口」、「緯度」、「首都」、「旧首都」、「最南端都市」・・・）に増加する。

【０３２９】

類似キーワード決定部４１が、第４のキーワードと類似する類似キーワードを決定する（ステップＳ３５）。例えば、第４のキーワード「旧首都」に類似する類似キーワードとして、キーワード入力部１１に入力された第２のキーワード（と同一の第４のキーワード）である「首都」が決定される。

【０３３０】

キーワード入力部１１へのキーワードの入力がある間（ステップＳ３６）は、上述したステップＳ３１～ステップＳ３５の処理が繰り返される。

【０３３１】

ステップＳ３６において、キーワード入力部１１への入力キーワードの入力がなくなると、質問作成部４２が、第３のキーワードと第４のキーワードとにより構成される質問を作成する（ステップＳ３７）。例えば、「日本の旧首都は？」、「アメリカの面積は？」、「ドイツの緯度は？」・・・といった質問を作成する。質問応答装置４が質問作成部４２を備えない構成を採るときは、上記ステップＳ３７の処理は、省略される。

【０３３２】

解答候補抽出部４３は、作成された各質問に対する解答候補を抽出する（ステップＳ３８）。すなわち、解答候補抽出部４３は、知識ＤＢ２５から、質問作成部４２によって作成された各質問を構成する第３のキーワードと第４キーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、各質問を構成する第４のキーワードが類似する類似キーワードに対応付けられた解答タイプに適合する言語表現を、解答候補として抽出する。

【０３３３】

ステップＳ３８において、解答候補抽出部４３は、質問応答装置４が質問作成部４２を省略する構成を採るときは、知識ＤＢ２５から、キーワード増加部１２によって出力される第３のキーワードと第４のキーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、第４のキーワードに類似する類似キーワードに対応付けられて解答タイプ入力部３１に入力された解答タイプに適合する言語表現を、第３のキーワードと当該第４のキーワードとによって構成される各質問に対する解答候補として抽出する。

【０３３４】

例えば、第４のキーワード「緯度」によって構成される質問に対する解答候補の抽出には、第４のキーワード「緯度」が類似する類似キーワード「面積」に対応付けられた解答タイプ「固有名詞（数値）」を用いる。

【０３３５】

また、例えば、第４のキーワード「旧首都」によって構成される質問に対する解答候補の抽出には、第４のキーワード「旧首都」が類似する類似キーワードに対応付けられた解答タイプ「固有名詞（地名）」を用いる。

【０３３６】

そして、解答表出力部１５が、解答表を出力する（ステップＳ３９）。例えば図１９に示すような解答表が出力される。

【０３３７】

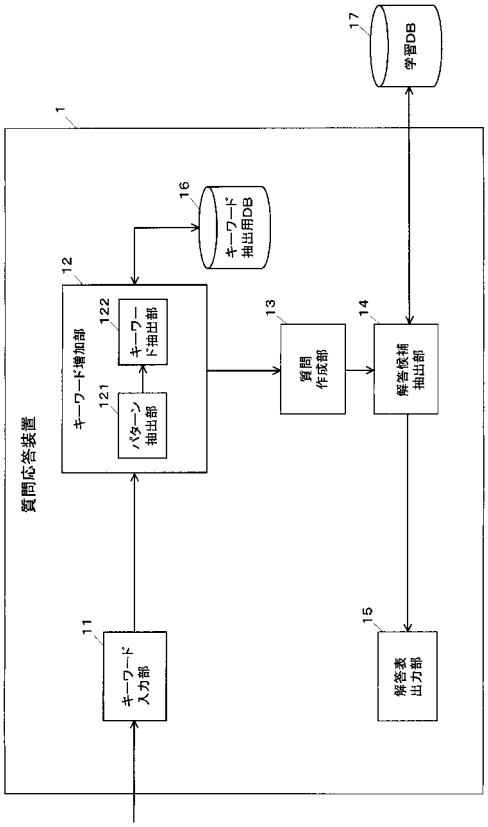
図２０は、本発明の第４の実施の形態における質問応答装置の構成の別の例を示す図である。質問応答装置４０は、入力されたキーワードを増加し、増加したキーワードにより構成される質問に対する解答を出力する装置である。図２０中に示す質問応答装置４０が備える構成要素のうち、図１に示す質問応答装置１が備える構成要素または図６に示す質問応答装置１０または図１７に示す質問応答装置４が備える構成要素と同一の符号が付けられたものは、当該質問応答装置１または質問応答装置１０または質問応答装置４が備える構成要素と同様の機能を有する。本発明の実施の形態においては、図２０に示す構成から質問作成部４２を省略し、解答候補抽出部４３が、キーワード増加部６０によって出力された第３のキーワードと第４のキーワードとによって構成される質問に対する解答候補を抽出し、出力する構成を採ってもよい。

【０３３８】
上記の質問応答装置４０を用いた場合の質問応答処理フローは、図１８に示す質問応答処理フローと、ステップＳ３３、ステップＳ３４の処理が異なる以外は、同様である。質問応答装置４０を用いた場合の質問応答処理フローの一例においては、図１８のステップＳ３３およびステップＳ３４の代わりに、キーワード増加部６０のキーワード抽出部１０１で、キーワード入力部１１に入力された第１のキーワードと同じ分野の単語を単語データＤＢ６１中から抽出し、第３のキーワードとして出力する。また、キーワード抽出部１０１で、キーワード入力部１１に入力された第２のキーワードと同じ分野の単語を単語データＤＢ６１中から抽出し、第４のキーワードとして出力する。
【０３３９】
また、質問応答装置４０を用いた場合の質問応答処理フローの別の例においては、図１８のステップＳ３３およびステップＳ３４の代わりに、キーワード増加部６０の類似度算出部１００が、キーワード入力部１１に入力された第１のキーワードとシソーラスＤＢ６２中の単語との類似度を算出し、キーワード増加部６０のキーワード抽出部１０１が、算出された類似度が予め定めた閾値以上の単語を第３のキーワードとして出力する。また、類似度算出部１００が、キーワード入力部１１に入力された第２のキーワードとシソーラスＤＢ６２中の単語との類似度を算出し、キーワード抽出部１０１が、算出された類似度が予め定めた閾値以上の単語を第４のキーワードとして出力する。
【０３４０】
また、キーワード抽出部１０１は、例えば、上記算出された類似度が大きい順に所定の個数の単語をシソーラスデータ中から取り出して、上記の第３のキーワード、第４のキーワードとして出力する構成を採ることもできる。
【０３４１】
また、本発明の第４の実施の形態においては、図１７に示す質問応答装置４または図２０に示す質問応答装置４０は、例えば、解答タイプ入力部３１に替えて、キーワード入力部１１に入力された第２のキーワードに対応付けられた疑問代名詞が入力される疑問代名詞入力部（図示を省略）と、上記疑問代名詞入力部に入力された疑問代名詞に基づいて、キーワード増加部１２（またはキーワード増加部６０）によって出力される出力キーワードによって構成される質問に対する解答の候補の言語表現の類型である解答タイプを推定する解答タイプ推定部（図示を省略）とを備える構成を採ってもよい。
【０３４２】
上記の構成においては、解答候補抽出部４３は、知識ＤＢ２５から、キーワード増加部１２（またはキーワード増加部６０）によって出力される第３のキーワードと第４のキーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、第４のキーワードに類似する類似キーワードに対応付けられて上記疑問代名詞入力部に入力された疑問代名詞に基づいて上記解答タイプ推定部が推定した解答タイプに適合する言語表現を、第３のキーワードと当該第４のキーワードとによって構成される各質問に対する解答候補として抽出してもよい。また、解答候補抽出部４３は、知識ＤＢ２５から、キーワード増加部１２（またはキーワード増加部６０）によって出力される第３のキーワードと第４のキーワードを含む文書を検索し、この検索処理で抽出された文書に含まれる言語表現のうち、第４のキーワードに類似する類似キーワードに対応付けられた疑問代名詞（すなわち、キーワード入力部１１に入力された第２のキーワードに対応付けされるものとして予め定められた疑問代名詞）に基づいて上記解答タイプ推定部が推定した解答タイプに適合する言語表現を、第３のキーワードと当該第４のキーワードとによって構成される各質問に対する解答候補として抽出してもよい。
【０３４３】
なお、本発明は、コンピュータにより読み取られ実行されるプログラムとして実施することもできる。本発明を実現するプログラムは、コンピュータが読み取り可能な、可搬媒体メモリ、半導体メモリ、ハードディスクなどの適当な記録媒体に格納することができ、これらの記録媒体に記録して提供され、または、通信インタフェースを介してネットワー

クを利用した送受信により提供されるものである。
【図面の簡単な説明】
【０３４４】
【図１】本発明の第１の実施の形態における質問応答装置の構成の一例を示す図である。
【図２】解答表の一例である。
【図３】キーワードの抽出結果に対する適合率・再現率の一例を示す図である。
【図４】正解データの一例を示す図である。
【図５】本発明の第１の実施の形態における質問応答処理フローの一例を示す図である。
【図６】本発明の第１の実施の形態における質問応答装置の構成の別の例を示す図である。
【図７】単語データＤＢのデータ構成例を示す図である。
【図８】シソーラスＤＢのデータ構成例を示す図である。
【図９】本発明の第２の実施の形態における質問応答装置の構成の一例を示す図である。
【図１０】解答表の一例である。
【図１１】本発明の第２の実施の形態における質問応答処理フローの一例を示す図である。
【図１２】本発明の第２の実施の形態の変形例１の構成例を示す図である。
【図１３】本発明の第３の実施の形態における質問応答装置の構成の一例を示す図である。
【図１４】本発明の第３の実施の形態における質問応答処理フローの一例を示す図である。
【図１５】解答表の一例である。
【図１６】本発明の第３の実施の形態の変形例１の構成例を示す図である。
【図１７】本発明の第４の実施の形態における質問応答装置の構成の一例を示す図である。
【図１８】本発明の第４の実施の形態における質問応答処理フローの一例を示す図である。
【図１９】解答表の一例である。
【図２０】本発明の第４の実施の形態における質問応答装置の構成の別の例を示す図である。
【図２１】解の候補と得点のリストの例である。
【図２２】解の候補の得点を単純に加算する方法を用いた出力結果の例である。
【図２３】質問に対する出力結果の例である。
【図２４】質問に対する出力結果の例である。
【図２５】質問に対する出力結果の例である。
【図２６】質問に対する出力結果の例である。
【図２７】サポートベクトルマシン法のマージン最大化の概念を示す図である。
【符号の説明】
【０３４５】
１、２、３、４、１０、２０、３０、４０　質問応答装置
１１　キーワード入力部
１２、１８、６０、６３　キーワード増加部
１３、２３、４２　質問作成部
１４、２４、４３　解答候補抽出部
１５　解答表出力部
１６　キーワード抽出用ＤＢ
１７　学習ＤＢ
２１　疑問代名詞入力部
２２　解答タイプ推定部
２５　知識ＤＢ

３１　解答タイプ入力部
４１　類似キーワード決定部
６１　単語データＤＢ
６２　シソーラスＤＢ
１００　類似度算出部
１０１、１２２　キーワード抽出部
１２１　パターン抽出部

【図１】



【図２】

	面積	人口	緯度
日本	A1	A2	A3
アメリカ	B1	B2	B3
ドイツ	C1	C2	C3

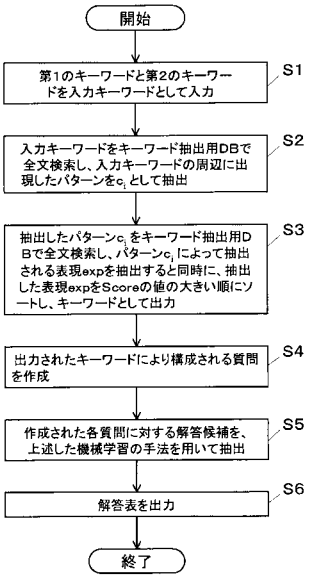
【図３】

	字種とKRを利用せず			字種の利用			KRの利用			字種とKRの利用		
	AP	RP	TP	AP	RP	TP	AP	RP	TP	AP	RP	TP
手法1	0.102	0.170	0.305	0.142	0.220	0.365	0.096	0.161	0.255	0.154	0.231	0.360
手法2	0.174	0.235	0.445	0.182	0.244	0.475	0.177	0.235	0.465	0.185	0.247	0.490
手法3	0.171	0.234	0.435	0.178	0.242	0.460	0.182	0.239	0.475	0.189	0.251	0.490
手法4	0.172	0.234	0.435	0.179	0.244	0.465	0.172	0.235	0.435	0.179	0.245	0.465
手法5	0.174	0.236	0.410	0.192	0.264	0.450	0.190	0.246	0.460	0.206	0.272	0.490

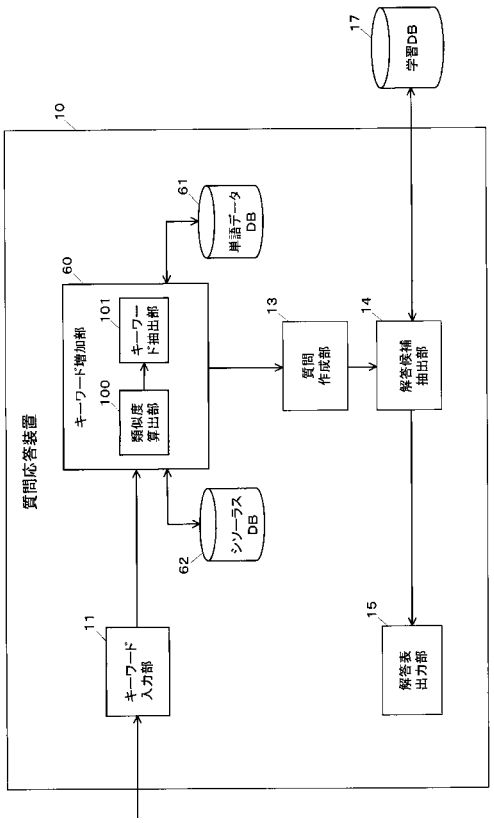
【図４】

アイスランド アイスランド共和国 ＩＳＬ
アイルランド アイルランド共和国 ＩＲＬ
アゼルバイジャン アゼルバイジャン共和国 ＡＺＥ
アゾレス諸島
アドゥイゲ アドゥイゲ共和国
アフガニスタン アフガニスタン共和国
アメリカ アメリカ合衆国 米国 米 ＵＳＡ
…

【図５】



【図６】



【図７】

分野	単語
国名	日本、アメリカ、ドイツ、・・・
数値表現	面積、人口、緯度、・・・
⋮	⋮

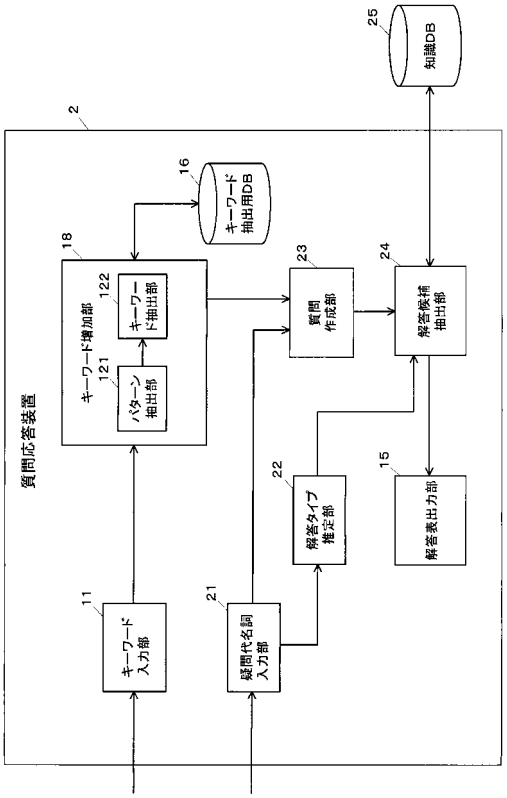
61

【図８】

単語	分類番号
日本	1259001012
アメリカ	1259004192
面積	1263013015
人口	1263010012
緯度	1263010015
⋮	⋮

62

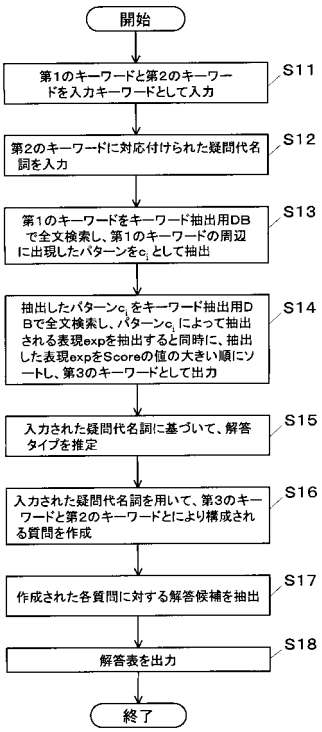
【図９】



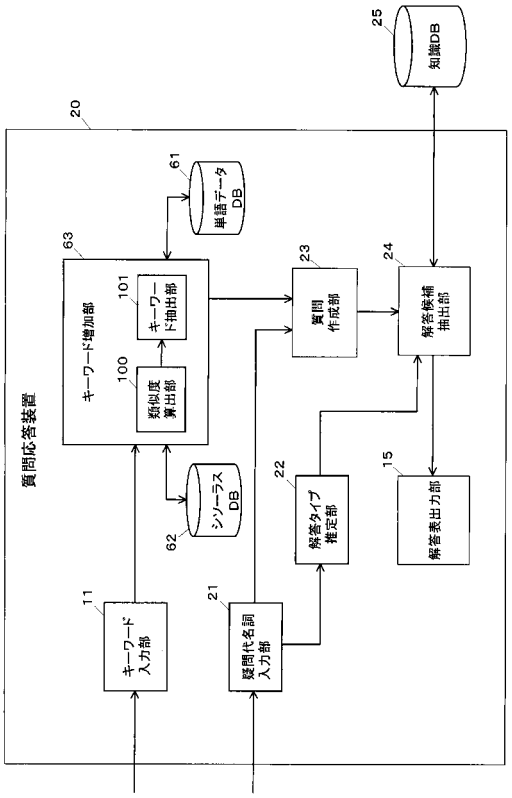
【図１０】

	首都
日本	東京
アメリカ	ワシントン
ドイツ	ベルリン
⋮	⋮

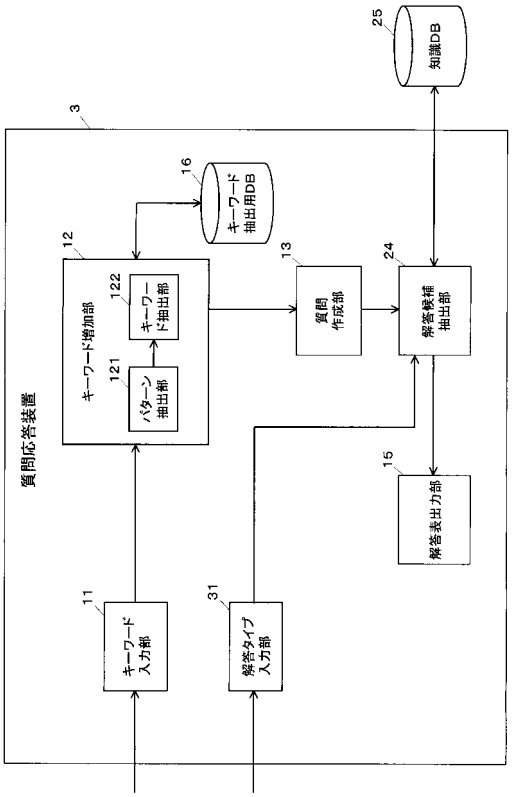
【図１１】



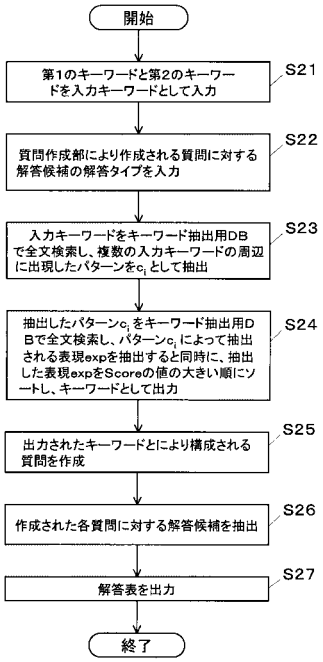
【図１２】



【図１３】



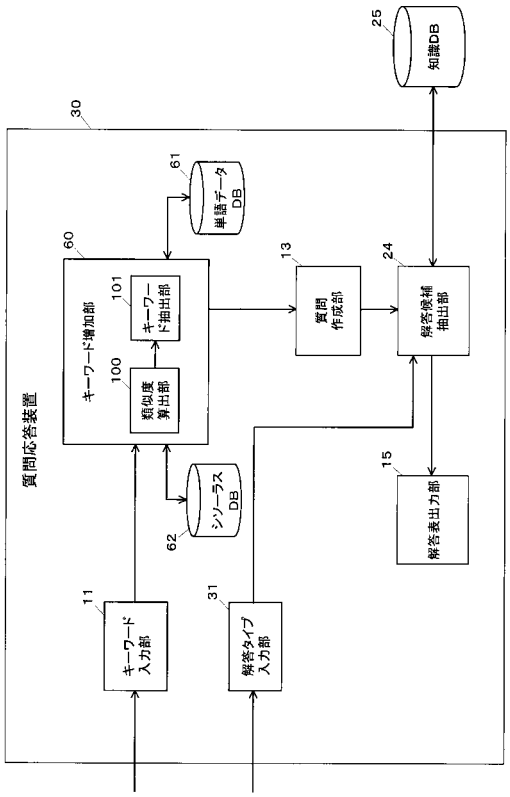
【図１４】



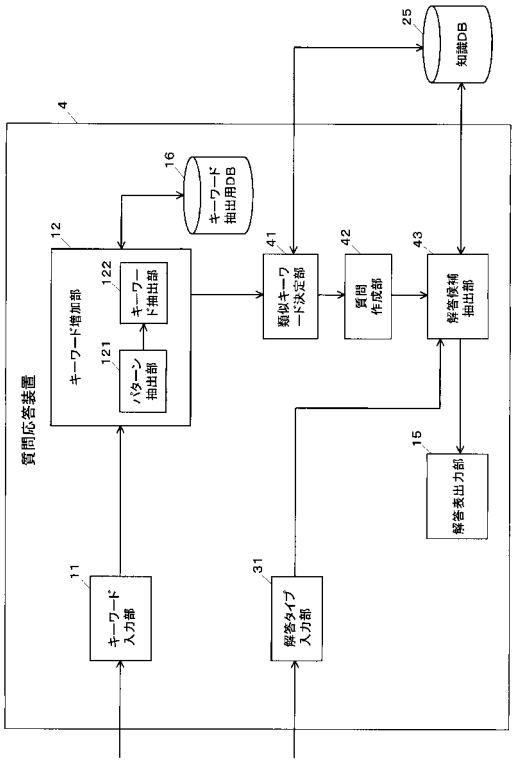
【図１５】

	首都	旧首都	最南端都市
日本	東京	江戸	a
アメリカ	ワシントン	フィラデルフィア	b
ドイツ	ベルリン	ボン	c
⋮	⋮	⋮	⋮

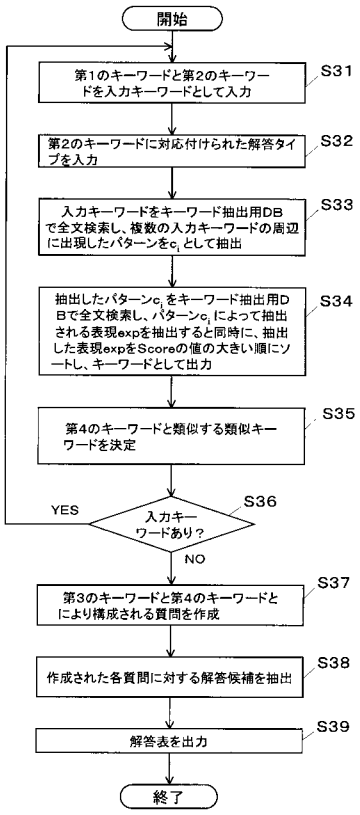
【図１６】



【図１７】



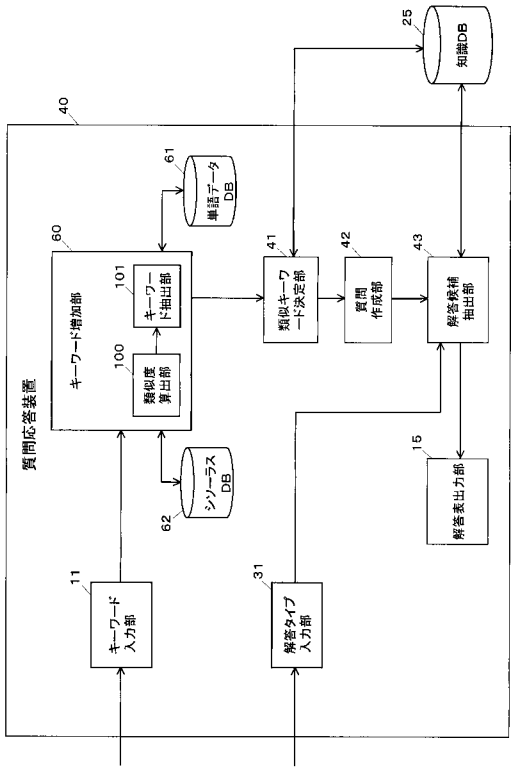
【図１８】



【図１９】

	面積	人口	緯度	首都	旧首都	最南端都市
日本	A1	A2	A3	東京	江戸	a
アメリカ	B1	B2	B3	ワシントン	フィラデルフィア	b
ドイツ	C1	C2	C3	ベルリン	—	c
イタリア	D1	D2	D3	ローマ	トリノ	d
フランス	E1	E2	E3	パリ	—	e
イギリス	FC1	F2	F3	ロンドン	—	f

【図２０】



【図２１】

順位	解の候補	得点	記事番号
1	京都	3.3	926324
2	東京	3.2	259312
3	東京	2.8	451245
4	東京	2.5	371922
5	東京	2.4	221328
6	北京	2.3	113127
…	…	…	…

【図２２】

順位	解の候補	得点	記事番号
1	東京	10.9	259312, 451245, 371922, 221328
2	京都	3.3	926324
3	北京	2.3	113127
…	…	…	…

【図２３】

順位	解の候補	得点	記事番号
1	京都	5.4	926324
2	東京	2.1	259312
3	東京	1.8	451245
4	東京	1.5	371922
5	東京	1.4	221328
6	北京	1.3	113127
…	…	…	…

【図２４】

順位	解の候補	得点	記事番号
1	東京	6.8	259312, 451245, 371922, 221328
2	京都	5.4	926324
3	北京	1.3	113127
…	…	…	…

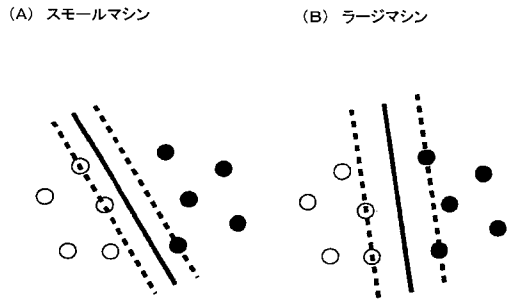
【図２５】

順位	解の候補	得点	記事番号
1	京都	5.4	926324
2	東京	2.8	259312, 451245, 371922, 221328
3	北京	1.3	113127
…	…	…	…

【図２６】

順位	解の候補	得点	記事番号
1	東京	4.3	259312, 451245, 371922, 221328
2	京都	3.3	926324
3	北京	2.3	113127
…	…	…	…

【図２７】



フロントページの続き

(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

審査官　鈴木　和樹

(56)参考文献　特開２００２－２２２１４７（ＪＰ，Ａ）
特開２００５－１５７５２４（ＪＰ，Ａ）
村田真樹、外２名，質問応答システムを用いた情報抽出，言語処理学会第６回年次大会ワークショップ論文集，日本，言語処理学会，２０００年　3月１０日，ｐ．３３－４０

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０