

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2006-293830
(P2006-293830A)

(43) 公開日 平成18年10月26日(2006.10.26)

(51) Int. Cl.	F I	テーマコード (参考)
G06F 17/30 (2006.01)	G06F 17/30 180A	5B075
G06F 17/28 (2006.01)	G06F 17/30 330C	5B091
	G06F 17/28 Z	

審査請求 未請求 請求項の数 15 O L (全 27 頁)

(21) 出願番号	特願2005-115893 (P2005-115893)	(71) 出願人	000005496 富士ゼロックス株式会社 東京都港区赤坂二丁目17番22号
(22) 出願日	平成17年4月13日 (2005.4.13)	(74) 代理人	100086531 弁理士 澤田 俊夫
		(74) 代理人	100093241 弁理士 宮田 正昭
		(74) 代理人	100101801 弁理士 山田 英治
		(72) 発明者	杉原 大悟 神奈川県足柄上郡中井町境430 グリーン テクナかい 富士ゼロックス株式会社内
		(72) 発明者	増市 博 神奈川県足柄上郡中井町境430 グリーン テクナかい 富士ゼロックス株式会社内 最終頁に続く

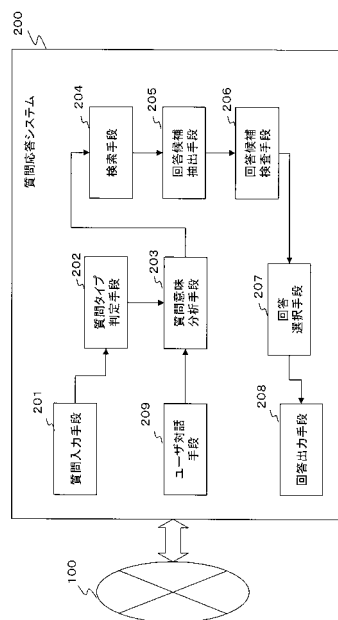
(54) 【発明の名称】 質問応答システム、およびデータ検索方法、並びにコンピュータ・プログラム

(57) 【要約】

【課題】 質問応答システムにおいて、質問に対して最適な回答を提供可能とした構成を実現する。

【解決手段】 入力質問に基づく検索処理結果に基づいて初期回答候補を抽出し、抽出した回答候補についての検査処理を実行し、質問に対する最適な回答を選択可能とした。回答候補を適用したクエリによって検索を行い、検索結果として得られた文の構成語が、入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する。本構成により、質問に対する最適な回答を回答候補として選択してクライアントに提供することが可能となる。

【選択図】 図2



【特許請求の範囲】

【請求項 1】

質問を入力する質問入力手段と、
入力質問に基づく検索処理を実行する検索手段と、
前記検索手段の検索結果に基づいて初期回答候補を抽出する回答候補抽出手段と、
前記回答候補抽出手段の抽出した回答候補についての検査処理を実行する回答候補検査手段と、
前記回答候補検査手段によって最終的に選択された回答を出力する回答出力手段を有し

、
前記回答候補検査手段は、

前記回答候補抽出手段の抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語が、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行する構成であることを特徴とする質問応答システム。

10

【請求項 2】

前記回答候補検査手段は、

前記入力質問から抽出される Q F (質問文の主題) と、前記回答候補抽出手段の抽出した回答候補との組み合わせからなるクエリを生成し、該クエリによる検索によってヒットした文を持つクエリに含まれる回答候補のみを選択する第 1 検査処理と、

前記第 1 検査処理において選択された回答候補について、前記第 1 検査処理において取得したヒット文の構成語について、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定する第 2 検査処理を実行する構成であり、該第 2 検査処理において、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行する構成であることを特徴とする請求項 1 に記載の質問応答システム。

20

【請求項 3】

前記回答候補検査手段は、

前記入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を修飾する語を抽出し、前記 Q F (質問文の主題) を修飾する語の語彙意味を解析する処理を実行する質問意味分析手段を備えることを特徴とする請求項 1 に記載の質問応答システム。

30

【請求項 4】

前記回答候補検査手段は、

前記質問意味分析手段が解析した前記入力質問における Q F (質問文の主題) を修飾する語の語彙意味と、前記回答候補抽出手段の抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語との比較処理を実行する構成であることを特徴とする請求項 3 に記載の質問応答システム。

【請求項 5】

前記質問意味分析手段は、

前記 Q F (質問文の主題) を修飾する語の語彙意味解析処理において、シソーラス辞書における前記 Q F (質問文の主題) と、前記 Q F (質問文の主題) を修飾する語の意味距離を解析し、より近い意味距離を持つ語彙意味を、前記 Q F (質問文の主題) を修飾する語の語彙意味として優先的に選択する処理を実行する構成であることを特徴とする請求項 3 に記載の質問応答システム。

40

【請求項 6】

前記質問意味分析手段は、

入力質問の構文意味解析処理を実行し、入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を修飾する語を抽出する処理を実行する構成であることを特徴とする請求項 3 に記載の質問応答システム。

【請求項 7】

前記質問意味分析手段は、

50

入力質問の構文意味解析時に格フレーム辞書を用いて、入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を直接修飾する動詞を含む語句部分を抽出する処理を実行する構成であることを特徴とする請求項 3 に記載の質問応答システム。

【請求項 8】

データ検索方法であり、
質問を入力する質問入力ステップと、
入力質問に基づく検索処理を実行する検索ステップと、
前記検索ステップにおける検索結果に基づいて初期回答候補を抽出する回答候補抽出ステップと、
前記回答候補抽出ステップにおいて抽出した回答候補についての検査処理を実行する回答候補検査ステップと、
前記回答候補検査ステップにおいて最終的に選択された回答を出力する回答出力ステップを有し、
前記回答候補検査ステップは、
前記回答候補抽出ステップにおいて抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語が、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行することを特徴とするデータ検索方法。

【請求項 9】

前記回答候補検査ステップは、
前記入力質問から抽出される Q F (質問文の主題) と、前記回答候補抽出ステップの抽出した回答候補との組み合わせからなるクエリを生成し、該クエリによる検索によってヒットした文を持つクエリに含まれる回答候補のみを選択する第 1 検査処理と、
前記第 1 検査処理において選択された回答候補について、前記第 1 検査処理において取得したヒット文の構成語について、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定する第 2 検査処理を実行し、第 2 検査処理において、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行することを特徴とする請求項 8 に記載のデータ検索方法。

【請求項 10】

前記回答候補検査ステップは、
入力質問の解析処理を実行し、入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を修飾する語を抽出し、
前記 Q F (質問文の主題) を修飾する語の語彙意味を解析する処理を実行する質問意味分析ステップを備えることを特徴とする請求項 8 に記載のデータ検索方法。

【請求項 11】

前記質問意味分析ステップは、
入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を修飾する語を抽出し、さらに、Q F (質問文の主題) を修飾する語についての語彙意味の解析処理を実行するステップであり、
前記回答候補検査ステップは、
前記質問意味分析ステップにおいて解析した前記入力質問における Q F (質問文の主題) を修飾する語の語彙意味と、前記回答候補抽出ステップの抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語との比較処理を実行するステップであることを特徴とする請求項 10 に記載のデータ検索方法。

【請求項 12】

前記質問意味分析ステップは、
前記 Q F (質問文の主題) を修飾する語の語彙意味解析処理において、シソーラス辞書における前記 Q F (質問文の主題) と、前記 Q F (質問文の主題) を修飾する語の意味距離を解析し、より近い意味距離を持つ語彙意味を、前記 Q F (質問文の主題) を修飾する語の語彙意味として優先的に選択する処理を実行することを特徴とする請求項 10 に記載

のデータ検索方法。

【請求項 13】

前記質問意味分析ステップは、

入力質問の構文意味解析処理を実行し、入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を修飾する語を抽出する処理を実行することを特徴とする請求項 10 に記載のデータ検索方法。

【請求項 14】

前記質問意味分析ステップは、

入力質問の構文意味解析時に格フレーム辞書を用いて、入力質問から Q F (質問文の主題) と、該 Q F (質問文の主題) を直接修飾する動詞を含む語句部分を抽出する処理を実行することを特徴とする請求項 10 に記載のデータ検索方法。 10

【請求項 15】

データ検索処理をコンピュータ上で実行させるコンピュータ・プログラムであり、質問を入力する質問入力ステップと、

入力質問に基づく検索処理を実行する検索ステップと、

前記検索ステップにおける検索結果に基づいて初期回答候補を抽出する回答候補抽出ステップと、

前記回答候補抽出ステップにおいて抽出した回答候補についての検査処理を実行する回答候補検査ステップと、

前記回答候補検査ステップにおいて最終的に選択された回答を出力する回答出力ステップを有し、 20

前記回答候補検査ステップは、

前記回答候補抽出ステップにおいて抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語が、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行するステップであることを特徴とするコンピュータ・プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、質問応答システム、およびデータ検索方法、並びにコンピュータ・プログラムに関する。さらに詳細には、質問文を入力して、質問に対応する回答を提供するシステムにおいて、質問に対して、よりの確な回答を提供可能とした質問応答システム、およびデータ検索方法、並びにコンピュータ・プログラムに関する。 30

【背景技術】

【0002】

昨今、インターネットなどを介したネットワーク通信が盛んになり、様々なサービスがネットワークを介して行なわれている。ネットワークを介したサービスの1つとして検索サービスがある。検索サービスは、例えばネットワーク接続されたパーソナルコンピュータ、携帯端末などのユーザ端末からの検索要求を検索サーバが受信し、検索サーバにおいて、検索要求に応じた処理を実行して処理結果をユーザ端末に送信するサービスである。 40

【0003】

例えば、インターネットを介した検索処理を実行する場合、検索サービスを提供している Web サイトにアクセスし、Web サイトの提示メニューに従って、キーワード、カテゴリなどの検索条件を入力しサーバに送信する。サーバは、これらの検索条件に従って、処理を実行し、処理結果をユーザ端末に表示する。

【0004】

データ検索処理には様々な態様がある。例えば、ユーザがキーワードを入力し、入力したキーワードを含む文書の一覧情報をユーザに提示するキーワードベースの検索方式や、ユーザが質問文を入力して質問に対する回答を提供するいわゆる質問応答システムなどが 50

ある。質問応答システムは、ユーザがキーワードを選択する必要がなく、また、質問に対する回答のみを受領することができるシステムであり利用が広まっている。

【0005】

多くの質問応答システムは、質問に対する回答候補を、例えば、様々なWebページや、アクセス可能なデータベースのように組織化されていない文書集合、いわゆるオープンドメイン文書集合から抽出する処理を実行する。このような回答候補抽出を行う質問応答システムにおいて、検索によって得られた回答候補が、クライアントからの質問に対する回答として適切な回答であるか否かを検証する技術について研究されている。

【0006】

例えば、非特許文献1には、回答候補の抽出に適用した検索対象テキスト集合以外のテキスト集合を用いた検索によって抽出した回答候補が、質問に対する回答として適切であるかを検証する技術を開示している。具体的には、クライアントからの質問文の主題と、検索によって得られた回答候補が、例えばシソーラス辞書において上位下位関係にあるか否か、あるいは、回答として数値を問う質問文である場合には、検索によって得られた回答候補が質問文の主題にあっていないか否かなどを確認する処理であり、質問の主題と、取得された回答候補との関係を表す判定パターンとコーパス(検索対象言語データ)を用いて質問に対する回答妥当性を検査する構成を開示している。特許文献1にも、数量表現についての検査、例えば、トラの数量を表すのは「頭」であるのが妥当であることの確認を行う処理構成について開示している。

10

【0007】

このように質問応答システムにおいて、オープンドメインな情報源(知識源ともいう)を利用して検索した回答候補の妥当性を判定する技術は、いくつか提案されている。しかし、このような回答候補の検査法は、基本的に以下のような手順を必要とする。

20

イ：人手によるパターンを用いて質問文からQF(質問文の主題)を抽出する。

例えば、質問文「国民栄誉賞を受賞した映画監督は誰ですか?」という質問からは、「映画監督」をQFとして取得する。

ロ：既存の質問応答システムと同様の手法にて、QFに基づく検索を実行し、回答候補を取得。例えば、「小淵恵三」と「黒沢明」を得る。

ハ：QF(質問文の主題)と回答候補とからなるパターンを生成し、生成パターンを検索文字列としてコーパス(検索対象言語データ)を検索する。例えば、「小淵恵三という映画監督」と「黒沢明という映画監督」といった文字列からなるパターンによって、コーパスを再度、検索し、ヒットする検索結果が得られた場合、このパターンに適用した回答候補が質問に対する妥当性が高いと判定し、この回答候補のみを質問者に対する回答として出力する。

30

【0008】

しかし、上記の回答検査手法では、QF(質問文の主題)と回答候補からなるパターンを生成して検索を行うものであり、クライアントからの質問文から取得するのは、QF(質問文の主題)のみであり、それ以上の情報を取得して検査に利用する構成とはなっていない。

【0009】

このような検査手法では、以下のような問題が生ずる可能性がある。例えば、以下のような質問を考える。質問「2003年に広島に行った野球選手は誰ですか?」

40

この質問に対するQF(質問文の主題)として取得される語は「野球選手」であり、検索の結果として得られた回答候補(A, B...)に対して、例えば、

[Aは野球選手である]

[Bは野球選手である]

等のパターンを生成して、コーパスの検索による検査を行う。

【0010】

しかし、この検査は不十分である場合がある。すなわち、質問「2003年に広島に行った野球選手は誰ですか?」

50

に含まれる「広島」は、質問文を入力したユーザの意図として「広島カーブ」という野球チームであった場合、「2003年、広島、選手」という検索キーワードを用いて得た回答候補には、「広島カーブの野球選手」以外の「選手」が含まれることは十分に考えられるからである。「2003年、広島、野球選手」というキーワードを元に得た回答候補には、広島カーブの対戦相手や選手のトレード先など他球団の選手名が含まれてしまうことが考えられる。

【0011】

この時、質問から抽出するQF（質問文の主題）として「野球選手」のみを抽出し、このQF（質問文の主題）と回答候補からなる文字列パターンを用いて検査をした場合には、他球団の選手である回答候補でも、コーパス中にヒット文が出現する可能性は十分あり、回答候補の検査がパスし、ユーザの質問に対する回答の妥当性ありという誤った判定がなされる。

10

【0012】

例えば、阪神の選手である「矢野」という選手が回答候補として得られた場合に、上述の手法によって、

質問に対するQF（質問文の主題）＝「野球選手」

回答候補＝「矢野」

を適用して、

「野球選手である矢野」という文字列パターンが生成され、

この文字列パターン「野球選手である矢野」をクエリとしてコーパスの検索を行なえば、コーパス中にヒット文が検出される可能性は、十分たかいものと考えられる。また、QFとして「広島に行った野球選手」という文字列を使用した場合にも、QF中の「広島」の意味が、「地名」なのか「スポーツチーム」なのかという語彙意味的な曖昧性が存在し、有効な検査を行い得ないことが考えられる。

20

【特許文献1】特開2004-118647号公報

【非特許文献1】「秋葉友良，藤井敦，伊藤克，"質問応答における常識的な選択と期待効用に基づく回答群の決定"、情報処理学会研究報告，2004-NL-163，p.131-138」

【発明の開示】

【発明が解決しようとする課題】

30

【0013】

上述したように、ユーザの質問に対応する回答を提供する質問応答システムにおける回答候補検査手法には、未だ不十分な点が多く、誤った回答を質問者（クライアント）に提示してしまう場合がある。本発明は、このような問題に鑑みてなされたものであり、既存の手法では回答候補の検査が不十分となる質問文に対しても有効な、より精度の高い検査により、質問に対する回答としてより適切な回答を選択することを可能とした質問応答システム、およびデータ検索方法、並びにコンピュータ・プログラムを提供することを目的とする。

【課題を解決するための手段】

【0014】

40

本発明の第1の側面は、質問を入力する質問入力手段と、入力質問に基づく検索処理を実行する検索手段と、前記検索手段の検索結果に基づいて初期回答候補を抽出する回答候補抽出手段と、前記回答候補抽出手段の抽出した回答候補についての検査処理を実行する回答候補検査手段と、前記回答候補検査手段によって最終的に選択された回答を出力する回答出力手段を有し、前記回答候補検査手段は、前記回答候補抽出手段の抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語が、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行する構成であることを特徴とする質問応答システムにある。

【0015】

50

さらに、本発明の質問応答システムの一実施態様において、前記回答候補検査手段は、前記入力質問から抽出されるQ F（質問文の主題）と、前記回答候補抽出手段の抽出した回答候補との組み合わせからなるクエリを生成し、該クエリによる検索によってヒットした文を持つクエリに含まれる回答候補のみを選択する第1検査処理と、

前記第1検査処理において選択された回答候補について、前記第1検査処理において取得したヒット文の構成語について、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定する第2検査処理を実行する構成であり、該第2検査処理において、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行する構成であることを特徴とする。

【0016】

10

さらに、本発明の質問応答システムの一実施態様において、前記回答候補検査手段は、前記入力質問からQ F（質問文の主題）と、該Q F（質問文の主題）を修飾する語を抽出し、前記Q F（質問文の主題）を修飾する語の語彙意味を解析する処理を実行する質問意味分析手段を備えることを特徴とする。

【0017】

さらに、本発明の質問応答システムの一実施態様において、前記回答候補検査手段は、前記質問意味分析手段が解析した前記入力質問におけるQ F（質問文の主題）を修飾する語の語彙意味と、前記回答候補抽出手段の抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語との比較処理を実行する構成であることを特徴とする。

【0018】

20

さらに、本発明の質問応答システムの一実施態様において、前記質問意味分析手段は、前記Q F（質問文の主題）を修飾する語の語彙意味解析処理において、シソーラス辞書における前記Q F（質問文の主題）と、前記Q F（質問文の主題）を修飾する語の意味距離を解析し、より近い意味距離を持つ語彙意味を、前記Q F（質問文の主題）を修飾する語の語彙意味として優先的に選択する処理を実行する構成であることを特徴とする。

【0019】

さらに、本発明の質問応答システムの一実施態様において、前記質問意味分析手段は、入力質問の構文意味解析処理を実行し、入力質問からQ F（質問文の主題）と、該Q F（質問文の主題）を修飾する語を抽出する処理を実行する構成であることを特徴とする。

【0020】

30

さらに、本発明の質問応答システムの一実施態様において、前記質問意味分析手段は、入力質問の構文意味解析時に格フレーム辞書を用いて、入力質問からQ F（質問文の主題）と、該Q F（質問文の主題）を直接修飾する動詞を含む語句部分を抽出する処理を実行する構成であることを特徴とする。

【0021】

さらに、本発明の第2の側面は、データ検索方法であり、質問を入力する質問入力ステップと、入力質問に基づく検索処理を実行する検索ステップと、前記検索ステップにおける検索結果に基づいて初期回答候補を抽出する回答候補抽出ステップと、前記回答候補抽出ステップにおいて抽出した回答候補についての検査処理を実行する回答候補検査ステップと、前記回答候補検査ステップにおいて最終的に選択された回答を出力する回答出力ステップを有し、前記回答候補検査ステップは、前記回答候補抽出ステップにおいて抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語が、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行することを特徴とするデータ検索方法にある。

40

【0022】

さらに、本発明のデータ検索方法の一実施態様において、前記回答候補検査ステップは、前記入力質問から抽出されるQ F（質問文の主題）と、前記回答候補抽出ステップの抽出した回答候補との組み合わせからなるクエリを生成し、該クエリによる検索によってヒットした文を持つクエリに含まれる回答候補のみを選択する第1検査処理と、前記第1検査

50

査処理において選択された回答候補について、前記第1検査処理において取得したヒット文の構成語について、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定する第2検査処理を実行し、第2検査処理において、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行することを特徴とする。

【0023】

さらに、本発明のデータ検索方法の一実施態様において、前記回答候補検査ステップは、入力質問の解析処理を実行し、入力質問からQF（質問文の主題）と、該QF（質問文の主題）を修飾する語を抽出し、前記QF（質問文の主題）を修飾する語の語彙意味を解析する処理を実行する質問意味分析ステップを備えることを特徴とする。

10

【0024】

さらに、本発明のデータ検索方法の一実施態様において、前記質問意味分析ステップは、入力質問からQF（質問文の主題）と、該QF（質問文の主題）を修飾する語を抽出し、さらに、QF（質問文の主題）を修飾する語についての語彙意味の解析処理を実行するステップであり、前記回答候補検査ステップは、前記質問意味分析ステップにおいて解析した前記入力質問におけるQF（質問文の主題）を修飾する語の語彙意味と、前記回答候補抽出ステップの抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語との比較処理を実行するステップであることを特徴とする。

【0025】

さらに、本発明のデータ検索方法の一実施態様において、前記質問意味分析ステップは、前記QF（質問文の主題）を修飾する語の語彙意味解析処理において、シソーラス辞書における前記QF（質問文の主題）と、前記QF（質問文の主題）を修飾する語の意味距離を解析し、より近い意味距離を持つ語彙意味を、前記QF（質問文の主題）を修飾する語の語彙意味として優先的に選択する処理を実行することを特徴とする。

20

【0026】

さらに、本発明のデータ検索方法の一実施態様において、前記質問意味分析ステップは、入力質問の構文意味解析処理を実行し、入力質問からQF（質問文の主題）と、該QF（質問文の主題）を修飾する語を抽出する処理を実行することを特徴とする。

【0027】

さらに、本発明のデータ検索方法の一実施態様において、前記質問意味分析ステップは、入力質問の構文意味解析時に格フレーム辞書を用いて、入力質問からQF（質問文の主題）と、該QF（質問文の主題）を直接修飾する動詞を含む語句部分を抽出する処理を実行することを特徴とする。

30

【0028】

さらに、本発明の第3の側面は、データ検索処理をコンピュータ上で実行させるコンピュータ・プログラムであり、質問を入力する質問入力ステップと、入力質問に基づく検索処理を実行する検索ステップと、前記検索ステップにおける検索結果に基づいて初期回答候補を抽出する回答候補抽出ステップと、前記回答候補抽出ステップにおいて抽出した回答候補についての検査処理を実行する回答候補検査ステップと、前記回答候補検査ステップにおいて最終的に選択された回答を出力する回答出力ステップを有し、前記回答候補検査ステップは、前記回答候補抽出ステップにおいて抽出した回答候補を適用したクエリによる検索結果として得られた文の構成語が、前記入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する処理を実行するステップであることを特徴とするコンピュータ・プログラムにある。

40

【0029】

なお、本発明のコンピュータ・プログラムは、例えば、様々なプログラム・コードを実行可能なコンピュータシステムに対して、コンピュータ可読な形式で提供する記憶媒体、通信媒体、例えば、CDやFD、MOなどの記録媒体、あるいは、ネットワークなどの通信媒体によって提供可能なコンピュータ・プログラムである。このようなプログラムをコ

50

ンピュータ可読な形式で提供することにより、コンピュータシステム上でプログラムに応じた処理が実現される。

【0030】

本発明のさらに他の目的、特徴や利点は、後述する本発明の実施例や添付する図面に基づくより詳細な説明によって明らかになるであろう。なお、本明細書においてシステムとは、複数の装置の論理的集合構成であり、各構成の装置が同一筐体内にあるものには限らない。

【発明の効果】

【0031】

本発明の構成によれば、質問に対する回答を提供するシステムにおいて、入力質問の語彙意味を解析し、入力質問の構成語の語彙意味に基づく回答候補検査処理を実行する構成としたので、質問に対する最適な回答を回答候補として選択してクライアントに提供することが可能となる。

10

【0032】

本発明の構成によれば、質問に対する回答を提供するシステムにおいて、入力質問に基づく検索処理結果に基づいて初期回答候補を抽出し、抽出した回答候補についての検査処理を実行し、質問に対する最適な回答を選択可能としたものであり、回答候補を適用したクエリによって検索を行い、検索結果として得られた文の構成語が、入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する回答候補検査処理を実行する構成としたので、質問に対する最適な回答を回答候補として選択してクライアントに提供することが可能となる。

20

【発明を実施するための最良の形態】

【0033】

以下、図面を参照しながら本発明の実施形態に係る質問応答システム、およびデータ検索方法、並びにコンピュータ・プログラムの詳細について説明する。

【0034】

[実施例1]

まず、図1を参照して、本発明の質問応答システムの利用形態の一例について説明する。図1は本発明の質問応答システム200をネットワークに接続したネットワーク構成を示す図である。図1に示すネットワーク100は、インターネットやイントラネットなどのネットワークであり、ネットワーク100には、質問応答システム200に対する質問を送信するユーザ端末としてのクライアント101-1~n、クライアント101-1~nに対する回答を取得するための素材としてのWebページを提供する様々なWebページ提供サーバ102A~Nやデータベース103a~nが接続されている。

30

【0035】

質問応答システム200は、クライアント101-1~nから、ユーザの生成した様々な質問文を入力し、入力質問に対する回答をクライアント101-1~nに提供する。質問に対する回答は、Webページ提供サーバ102A~Nの提供するWebページや、データベース103a~nに格納された文書データなどから取得する。Webページ提供サーバ102A~Nの提供するWebページや、データベース103a~nに格納されたデータが検索対象データでありコーパス、あるいは情報源、知識源などと呼ばれる。

40

【0036】

Webページ提供サーバ102A~Nは、WWW(World Wide Web)システムによる公開ページとしてWebページを提供する。Webページは、Webブラウザに表示されるデータ集合であり、テキストデータやHTMLによるレイアウト情報、文書中に埋め込まれた画像や音声、動画などによって構成される。Webページの集合はWebサイトであり、Webサイトは、トップページ(ホームページ)とトップページからリンクされた他のWebページで構成される。

【0037】

50

質問応答システム200の構成および処理について図2を参照して説明する。質問応答システム200は、ネットワーク100に接続され、ネットワーク100に接続されたクライアントからの質問を受領し、ネットワーク100に接続されたWebページ提供サーバの提供するWebページやその他のデータベースを情報源として回答を検索して、例えば回答候補からなるリストを生成してクライアントに提供する処理を実行する。

【0038】

本発明の質問応答システムは、クライアントからの質問に対する一次的な検索によって得られた回答候補の検査を行い、質問に対する回答候補の妥当性を判定する処理を実行する。本発明の質問応答システムでは、質問文を構成する語彙の意味を取得し、これを回答候補の検査に用いることで、より精度の高い回答候補検査を行うものである。

10

【0039】

具体的には、質問文中の語彙意味の組み合わせから質問文の構成要素の意味や関係を解析し、その解析結果に基づいて回答候補を選択する。例えば、

質問「2003年に広島に行った野球選手は誰ですか？」

という質問があった場合、

「野球選手が広島に行く」における「野球選手」と「広島」と「行く」との間に発生している関係は、「広島という場所に行く」という関係や「広島という球団に入団する」という関係であるということとを区別して取得する。そして、質問文の「広島」が「地名」で用いられているか「地名（あるいはスポーツチーム）」で用いられているかを判定し、その結果を用いて、質問応答システムの初期的な検索によって得られた回答候補の回答候補以外の語彙、上記の場合だと「野球選手が行った広島」の意味を判定することで、質問文の意図にあった文脈を判断し、それによって回答候補を峻別する。

20

【0040】

質問文中の語彙の語彙的な意味の関係をj用いて検査を行う場合の利点としては、以下のようjなものjが挙げられる。語彙的な意味をj用いるjことで表現のゆれを吸収しつつ、質問文のQF（質問文の主題）以外の情報を、回答候補の質問に対する妥当性の検査処理に組み込むことができ、検査の適合率を上げると同時に、回答候補を限定する精度を上げることができる。また、質問文中の語彙意味の組み合わせから、質問文の意図を細かく判断することができ、生じた質問の意味の曖昧性ごとに個別に対応することができる。このことで、特定の場合に対する外部データベースへのアクセスなどを実装すること、あるいはユーザに対する質問文の意図の聞き返しなど、質問文を構成する意味ごとの個別の対応が可能となる。

30

【0041】

本発明の質問応答システム200の構成について、図2を参照して説明する。質問応答システム200は、図2に示すように、質問タイプ判定手段201、回答候補抽出手段202、クエリ生成手段203、パッセージ検索手段204、形態素解析手段205、回答候補検査手段206、回答出力手段207を有する。以下、質問応答システム200の有する各処理手段の実行する処理について説明する。

【0042】

[質問入力手段]

質問入力手段201は、ネットワーク100を介して、クライアントからの質問文（入力質問）を入力する。以下では、具体的な質問例として、

（入力質問）「2003年に広島に行った野球選手は誰ですか？」

という質問がクライアントから入力されたと想定して、質問応答システム200における各手段が実行する処理について説明する。

40

【0043】

[質問タイプ判定手段]

質問タイプ判定手段202は、既存の多くの質問応答システムと同様に、入力された質問文の質問タイプを判定する。例えば、下記の質問文、

「2003年に広島に行った野球選手は誰ですか？」

50

からは、「誰」という文字列から「人物」についての質問タイプであると判定する。

【0044】

[質問意味分析手段]

質問意味分析手段203は、検索手段204における検索によって取得される回答候補の検査処理に適用するため、および、検索手段204における検索に適用する検索クエリを生成するため、質問文を構成する語彙の意味を取得する処理を実行する。

【0045】

質問意味分析手段203の実行する処理の具体例について説明する。まず、従来から知られる手法(例えば非特許文献1に記載の手法)を適用して、質問文からQF(質問文の主題)を取得する。

10

質問文「2003年に広島に行った野球選手は誰ですか？」

からは、QF(質問文の主題)として「野球選手」が選択される。

次に、取得された質問文のQF(野球選手)と、このQFを修飾する部分の語を検出する。質問文のQF(野球選手)を修飾する語の検出には、例えば構文意味解析処理が適用される。例えば、形態素解析、構文解析、意味解析、文脈解析などにより、質問文を解析し、質問文のQF(野球選手)を修飾する語を検出する。解析結果として質問文のQF(野球選手)に対しては、QFを修飾する語「広島」が検出される。

【0046】

次に、取得された質問文のQF(野球選手)を修飾する部分の語「広島」の語彙について、既存のシソーラス辞書を用いてその語彙意味を取得する。例えば、シソーラス辞書に基づいて、広島という語から「地名」と「球団」の二つの意味(語義)を得ることができる。

20

【0047】

得られた複数の語義に優先順位を与えたい場合には、二つの語義の類似性を与えるシソーラス上の距離を計測し、距離に基づいて優先順位を擬似的に与えることができる。例えば、「野球選手」に対して「広島」の二つの語義(「地名」と「球団」)のどちらがシソーラス上の距離として近いかを計測し、優先順位を与える。二つの語のシソーラス上の意味属性を x と y とし、 l_x と l_y をそれぞれの意味属性のシソーラスの根(ルート)からの階層の深さ、 L を二つの意味属性で一致している階層の深さとする。この時、二つの語義 x 、 y のシソーラス上の距離(意味距離)は一般的に以下で与えられる。

30

$$\text{dist}(x, y)$$

【0048】

「広島」の二つの意味「地名」と「球団」の意味属性をそれぞれ x_1 と x_2 とし、「野球選手」の意味属性を y とすると、

「広島(地名)」と「野球選手」との距離： $\text{dist}(x_1, y)$

「広島(球団)」と「野球選手」との距離： $\text{dist}(x_2, y)$

となり、 $\text{dist}(x_1, y)$ 、 $\text{dist}(x_2, y)$ を計算する。

この二つのシソーラス上の距離 $\text{dist}(x_1, y)$ 、 $\text{dist}(x_2, y)$ を比較して、より短い距離を持つ方を高い優先順位とする。以下では広島の意味として「球団 | (同義語)カーブ」が高い優先順位を得たとして説明を行う。

40

【0049】

ここで特記しておくべきは、本実施例では質問文に含まれるQF(質問文の主題)を用いて、質問文に含まれる「広島」の曖昧性を解決しているが、例えば、「2003年広島に行ったのは誰ですか？」のようにQF(質問文の主題)として特別な文字列を設定できないような場合には、「広島」の意味の曖昧性を解消せず、質問文中の語彙の組み合わせだけを取得したまま処理を継続し、回答候補に対する検査の段階で語彙意味の順位を判定することも考えられる。

【0050】

あるいは別の構成としては、質問文の語彙に複数の意味が存在する場合には、ユーザに対して該当する語彙に複数の意味がある旨を通知し、ユーザの返答をまって意味を決定し

50

てもよい。この処理は、図 2 に示すユーザ対話手段 209 を利用してユーザに判定を求める処理として実行される。

【0051】

質問文の構成語の曖昧性が解消しない場合には、後の回答候補検査手段 206 における回答候補検査処理の際に、語彙意味ごとに用意した制約を使い分けて回答候補の検査を行い、得られた検査結果を意味ごとに分けて出力する。このように、質問文中の語彙間の意味を用いて検査を行う手法では、適合率を落とさずに、精度向上のための個別対応が可能となる。

【0052】

[検索手段 204]

検索手段 204 は、既存の多くの質問応答システムと同様に、質問文から得られたキーワードを元に検索対象から文書を検索し、回答候補がありそうな文データ部分（パッセージ）を抽出する。ここでの検索対象データは、図 1 に示す Web ページ提供サーバ 102 A ~ N の提供する Web ページや、データベース 103 a ~ n に格納された文書データなどであり、コーパス、あるいは情報源、知識源などと呼ばれる。

10

【0053】

このデータ検索処理に関する具体的な技術については、例えば、「Isozaki, H., "NTT's Question Answering System for NTCIR QAC2", Working Notes of NTCIR-4 Workshop, pp. 326-332 (2004)」に示されている。

20

【0054】

具体的には、

質問文「2003年に広島に行った野球選手は誰ですか？」

から取得される「2003年、広島、野球選手」といったキーワードを用いて検索対象から文書を得、得られた文書から、回答候補が含まれていそうな文データ部分（パッセージ）を得る。

【0055】

[回答候補抽出手段]

回答候補抽出手段 205 は、既存の質問応答技術を用いて、検索手段 204 における検索処理によって取得したパッセージから、前述の質問タイプ判定手段 202 の判定した質問タイプ「人物」に相当する語を選択し、この選択語を回答候補として取得する。

30

【0056】

例えば、

質問文「2003年に広島に行った野球選手は誰ですか？」

から取得されるキーワード「2003年、広島、野球選手」の近傍に出現し、かつ語彙を示す NE (named entity) タグで「人物」タグが付与されている名詞をパッセージ中から取得して、回答候補とする。

【0057】

[回答候補検査手段]

回答候補検査手段 206 は、回答候補抽出手段 205 がパッセージから抽出した回答候補に対する検査を実行する。検査は、QF（質問文の主題）の文字列と、質問意味分析手段 203 で得られた語彙意味についての制約を用いて行う。例えば、

40

質問文「2003年に広島に行った野球選手は誰ですか？」

から取得されるキーワード「2003年、広島、野球選手」に基づく検索によって得られた回答候補中に、

「黒田」「矢野」「渡辺オーナー」という回答候補が存在した場合に、これを以下のような手順で検査する。

【0058】

（第 1 ステップ）

まず、既存の手法による QF（質問文の主題）の文字列を用いた検査を行う。

50

すなわち、

[Q F (質問文の主題) + 回答候補]

の文字列パターンからなる検索クエリを生成して、コーパス (情報源) に対する検索を行う。

【 0 0 5 9 】

具体的には、例えば W e b ページ提供サーバの提供する W e b ページなどの外部コーパスに対して、

- a) 「黒田という野球選手」
- b) 「矢野という野球選手」
- c) 「渡辺オーナーという野球選手」

などの検索クエリを作成し、

これらの各検索クエリに対応するヒット文書を保持する。

【 0 0 6 0 】

この処理によって、

- a) 「黒田という野球選手」
- b) 「矢野という野球選手」

の2つの検索クエリに対するヒット文書は検出されるが、

- c) 「渡辺オーナーという野球選手」

という検索クエリに対するヒット文書は検出されず、結果として、この検査において、回答候補として、野球選手である「黒田」や「矢野」が検査にパスし、回答候補として残されるが、「渡辺オーナー」は野球選手ではなく、ヒット文書が検出されず、検査にパスせず、回答候補から除外される。

【 0 0 6 1 】

(第 2 ステップ)、

次に、回答候補検査手段 2 0 6 は、第 1 ステップの検索において、各回答候補が得られた文書の文脈において用いられている検索キーワードが、質問意味分析手段 2 0 3 で得られた質問文の語彙意味の用法で用いられているかどうか、各回答候補について検査する。

【 0 0 6 2 】

この検査処理においては、回答候補抽出手段 2 0 5 で得た回答候補ごとに、その回答候補が得られた文書の文脈において用いられている検索キーワードが、質問意味解析手段 2 0 3 で得られた語彙意味で用いられているか否かを判定する。ここでは、検索キーワードとして、「 2 0 0 3 年、広島、野球選手」を適用し、

上記のステップ 1 の検査において、

- a) 「黒田という野球選手」
- b) 「矢野という野球選手」

の2つの検索クエリに対するヒット文書として検出されているので、これらの文書に含まれる文書中の「広島」が、質問意味解析手段 2 0 3 で得られた語彙意味で用いられているか否かを判定する。

【 0 0 6 3 】

質問意味解析手段 2 0 3 では、

質問文「 2 0 0 3 年に広島に行った野球選手は誰ですか？」

から取得された質問文の Q F (野球選手) を修飾する部分、すなわち「広島」の語彙について、既存のシソーラス辞書から得られた二つの意味「地名」と「球団」について、シソーラス上の距離 (意味距離) の計測、すなわち、

「広島 (地名) 」と「野球選手」との距離 : $d i s t (x 1 , y)$

「広島 (球団) 」と「野球選手」との距離 : $d i s t (x 2 , y)$

に基づいて、質問中の「広島」が「球団」であると判定している。

【 0 0 6 4 】

従って、上記のステップ 1 の検査において、

- a) 「黒田という野球選手」

10

20

30

40

50

b) 「矢野という野球選手」

の2つの検索クエリに対するヒット文書として検出された文書に含まれる文書中の「広島」が「球団」として適用されているかどうかを判定する。

【0065】

この判定処理のために、「矢野」と「黒田」が含まれるヒット文書に対して、それぞれ形態素解析処理を行う。図3に、前述のステップ1の検索処理、すなわち、

a) 「黒田という野球選手」

b) 「矢野という野球選手」

の2つの検索クエリに基づく検索におけるヒット文書に含まれる構成分の形態素解析結果を示す。形態素解析は、文を意味的最小単位である形態素(m o r p h e m e)に分節して品詞の認定処理を行う解析処理である。 10

【0066】

図3には、

A. 「矢野は広島市に行った。」に対する形態素解析結果の一部データ、

B. 「黒田は広島カーブに行った。」に対する形態素解析結果の一部データ、

これら2つのヒット文の形態素解析結果を示している。

【0067】

A. 「矢野は広島市に行った。」の「広島」に対する形態素解析結果は「名詞 - 固有名詞 - 地名 - 一般」であり、

B. 「黒田は広島カーブに行った。」の「広島」は「名詞 - 固有名詞 - 組織」である。二つの形態素解析結果において異なっているのは「地名」と「組織」である(「一般」部分は処理上で意味がないので除去する)。 20

【0068】

前述したように、質問意味分析手段203における

質問文「2003年に広島に行った野球選手は誰ですか？」

に対する語彙解析によって、質問文における「広島」が「球団」、すなわち組織であると解析されている。

B. 「黒田は広島カーブに行った。」の「広島」は、

質問意味分析手段203が解析した質問文の[広島(組織)]の持つ語彙意味と同様の用法で用いられているが、 30

A. 「矢野は広島市に行った。」の「広島」は、

質問意味分析手段203が解析した質問文の[広島(地名)]の持つ語彙意味と異なる用法で用いられている。

【0069】

従って、回答候補検査手段206は、

前述のステップ1の検索処理で選択された文書、すなわち、

A. 「矢野は広島市に行った。」、

B. 「黒田は広島カーブに行った。」、

から、B. 「黒田は広島カーブに行った。」に含まれる回答候補[黒田]を選択、あるいはスコア加算を行う。 40

【0070】

この回答候補検査処理によって、

質問文「2003年に広島に行った野球選手は誰ですか？」

に対する回答候補として、

「黒田」

のみを選択、あるいは最高スコアとした回答候補リストを生成することができる。

【0071】

なお、回答候補に対するスコア加算処理においては、シソーラスの意味属性間の距離と、その意味属性と関連があるかの正解との対応関係を学習し、十分な量の学習データによって学習されたスレッシュホールドによって、回答候補が含まれるヒット文書中の「広島 50

」の意味を判定し、質問の意図にあった回答候補を取得する構成とする。また、シソーラス辞書によっては、広島「球団」の同義語として「カーブ」という語彙が得られることがある。スポーツチームなどでは、「広島カーブ、野球選手」という検索クエリを用いて、直接選手のデータベースに問い合わせを行うプログラムを保持することも有効な方法である。

【0072】

[回答選択手段]

回答選択手段207は、回答候補検査手段206で選択された回答候補を、例えばスコア値に基づくランキングリストとして、回答候補リストを生成する。

【0073】

[回答出力手段]

回答出力手段208は、回答選択手段207において最終的に決定した回答候補をクライアントに対して出力する。

【0074】

これらの処理によって、

例えば、質問Q、すなわち、

質問Q：「2003年に広島に行った野球選手は誰ですか？」

に対する回答として、

「黒田」

のみを選択、あるいは最高スコアとした回答候補リストをクライアントに提示することができる。 20

【0075】

なお、質問意味分析手段203における分析処理の結果として、質問文におけるQF（質問文の主題）の修飾部分の語に関する語彙が複数存在し、曖昧性を解消しない場合には、QFとの意味的な距離、およびコーパス中のQFとの共起頻度の高い意味を優先的に出力するなどの工夫を行う構成としてもよい。

【0076】

例えば、「広島」と「野球選手」の共起頻度をカウントする際に、「広島」が「組織」というNEを持つ場合の頻度が、「地名」のNEを持つ場合よりも多い場合には、「広島カーブの黒田」のような、「組織」という意味で用いられる「広島」を含む文脈の回答候補を優先的に出力するなどの工夫が可能である。 30

【0077】

次に、図4のフローチャートを参照して、本発明の質問応答システムの実行する処理シーケンスについて説明する。

【0078】

ステップS101において、クライアントからの質問を入力すると、ステップS102において、入力された質問文の質問タイプを判定する。これは図2に示す質問タイプ判定手段202の実行する処理であり、

質問「2003年に広島に行った野球選手は誰ですか？」

からは、「誰」という文字列から「人物」についての質問タイプであると判定する。 40

【0079】

次に、ステップS103において、質問文を構成する語の語彙意味を取得する処理を実行する。これは、図2に示す質問意味分析手段203の処理であり、回答候補の検査処理に適用し、また、検索に適用する検索クエリを生成するために、質問文を構成する語彙の意味を取得する処理である。

質問「2003年に広島に行った野球選手は誰ですか？」

から、質問文のQF（野球選手）と、QFを修飾する語として「広島」が抽出され、QFを修飾する語「広島」について、シソーラス辞書を用いてその語彙意味を取得する。

【0080】

この場合、例えば、シソーラス辞書に基づいて、広島という語から「地名」と「球団」 50

の二つの意味（語義）を得ることができる。次に、ステップ S 1 0 4 において、シソーラス上の距離（意味距離）に基づいて、質問文の解析対象語の語彙意味を判定する。なお、このステップの処理は、Q F を修飾する語に複数の語彙意味がある場合に実行すればよい。

【 0 0 8 1 】

すなわち、上述の例では、「広島」について、シソーラス辞書から、「地名」と「球団」の二つの意味（語義）を得ることができるので、

質問に含まれる Q F（野球選手）と、各意味属性を持つ修飾語「広島（地名）」、「広島（球団）」との意味距離を測定する。すなわち、

「広島（地名）」と「野球選手」との距離： $dist(x_1, y)$

「広島（球団）」と「野球選手」との距離： $dist(x_2, y)$

の各距離を測定し、より近い意味距離を持つ方を、質問における Q F（野球選手）の修飾語「広島」の持つ語彙意味として選択する。

この場合、「広島」は、「球団」の語彙意味を持つと判定される。

【 0 0 8 2 】

次に、ステップ S 1 0 5 において、質問に対応する検索処理が実行される。この検索処理は、Web ページ、データベースなどのコーパス（情報源）を対象とした検索処理であり、図 2 に示す検索手段 2 0 4 が実行する。質問文から選択されたキーワードに基づいて生成されるクエリによる検索処理が実行される。

【 0 0 8 3 】

具体的には、例えば、

質問文「2003年に広島に行った野球選手は誰ですか？」

から取得される「2003年、広島、野球選手」といったキーワードを用いたクエリによって、検索対象から文書を得、得られた文書から、回答候補が含まれていそうな文データ部分（パッセージ）を取得する処理として実行される。

【 0 0 8 4 】

次に、ステップ S 1 0 6 において、検索処理によって取得したパッセージから、前述の質問タイプ判定手段の判定した質問タイプ、例えば「人物」に相当する語を選択し、この選択語を回答候補として取得する。この処理は、図 2 における回答候補抽出手段 2 0 5 の実行する処理である。

【 0 0 8 5 】

例えば、

質問文「2003年に広島に行った野球選手は誰ですか？」

から取得されるキーワード「2003年、広島、野球選手」に基づく検索によって得られたパッセージから、「人物」に相当する語として、「黒田」「矢野」「渡辺オーナー」という語を抽出し、これらを回答候補とする。

【 0 0 8 6 】

次に、ステップ S 1 0 7、ステップ S 1 0 8 において回答候補検査処理を実行する。これらの処理は、図 2 に示す回答候補検査手段 2 0 6 が実行する。

まず、ステップ S 1 0 7 において、

回答候補検査処理第 1 ステップを実行する。

これは、

[Q F（質問文の主題） + 回答候補]

の文字列パターンからなる検索クエリを生成して、コーパス（情報源）に対する検索を行う処理である。

【 0 0 8 7 】

具体的には、

質問文「2003年に広島に行った野球選手は誰ですか？」

から取得されるキーワード「2003年、広島、野球選手」に基づく検索によって得られた回答候補中に、「黒田」「矢野」「渡辺オーナー」という回答候補が存在した場合に

10

20

30

40

50

- a) 「黒田という野球選手」
- b) 「矢野という野球選手」
- c) 「渡辺オーナーという野球選手」

などの検索クエリを作成し、これらの各検索クエリに基づく検索を行い、ヒット文書を持つ回答候補のみを回答候補として維持し、ヒット文書を持たない回答候補は、回答候補から除外する。

【0088】

- a) 「黒田という野球選手」
- b) 「矢野という野球選手」

はヒット文書を持ち、「黒田」、「矢野」のみが回答候補として残る。

10

【0089】

ステップS108は、回答候補検査処理第2ステップであり、第1ステップ(S107)の検索において、各回答候補が得られた文書の文脈において用いられているワードが、質問意味分析手段203で得られた質問文の語彙意味の用法で用いられているかどうか、各回答候補について検査する処理である。

【0090】

ステップS107の回答候補検査処理第1ステップにおいて、

- a) 「黒田という野球選手」
- b) 「矢野という野球選手」

の2つの検索クエリに対するヒット文書として検出された文書、すなわち、

- A. 「矢野は広島市に行った。」
- B. 「黒田は広島カーブに行った。」

の各文書の形態素解析を実行し、質問意味分析手段203で得られた質問文の語彙意味の用法で用いられているかどうか、各回答候補について検査する。

20

【0091】

この場合、質問文

質問文「2003年に広島に行った野球選手は誰ですか？」

に対する語彙解析によって、質問文における「広島」が「球団」、すなわち組織であると解析されている。

30

B. 「黒田は広島カーブに行った。」の「広島」は、

質問意味分析手段203が解析した質問文の[広島(組織)]の持つ語彙意味と同様の用法で用いられているが、

A. 「矢野は広島市に行った。」の「広島」は、地名であり、質問文の[広島(組織)]の持つ語彙意味と異なる用法で用いられていることが判明する。

【0092】

この結果、B. 「黒田は広島カーブに行った。」に含まれる回答候補[黒田]が選択、あるいはスコアアップされる。

【0093】

ステップS109では、ステップS107, S108の回答候補検査処理によって選択された回答候補について、例えばスコア値に基づくランキングリストからなる回答候補リストを生成し、ステップS110において、クライアントに対して出力する。

40

【0094】

これらの処理によって、

例えば、質問Q、すなわち、

質問Q: 「2003年に広島に行った野球選手は誰ですか？」

に対する回答として、

「黒田」

のみを選択、あるいは最高スコアとした回答候補リストをクライアントに提示することができる。

50

【0095】

[その他の変更実施例]

次に、本発明に係る質問応答システムにおけるその他の実施例について説明する。

上述した実施例では、質問意味分析手段203は、質問文からQF(質問文の主題)を取得し、さらに、その修飾語を取得する処理を実行しているが、質問文の検査に格フレーム辞書、あるいは格フレーム辞書に相当するデータ構造を用いることが可能である。格フレームを制約として用いることで、格フレームの格要素から、動詞についてのクエリの拡張が可能となる。また、語彙間の意味のパターンについて格フレーム照合することで、表記のゆれの吸収を動詞にまで拡大して検査を行うことができる。

【0096】

また、「広島に行ったのは誰ですか?」のように明確にQFとなる文字列が得られない場合にも、格フレームの構造から主格と斜格の意味クラスのパターンを得ることができ、回答候補に対する検査が可能となる。

【0097】

以下に、質問意味分析手段202において、質問文の語彙間の意味を、格フレームを用いて判定し、それを用いて回答候補の検査を行う例を説明する。「2003年に広島に行った野球選手は誰ですか?」という質問がシステムに入力されたとして、説明する。

【0098】

まず、質問文から格フレームの構造を得るため、質問文を構文意味解析する処理を行う。最初に、質問文特有の語尾を除去する。この場合、質問文の語尾を除去し、質問文特有の語彙「誰」をダミーの文字列に変換する。

2003年に広島に行った野球選手は誰ですか?

2003年に広島に行った野球選手はAです。

得られた文に構文意味解析処理を施す。構文意味解析処理について説明する。日本語や英語を始めとする各種の言語で記述される自然言語は、本来抽象的でありまい性が高い性質を持つが、文章を数学的に取り扱うことにより、コンピュータ処理を行うことができる。この結果、機械翻訳や対話システム、検索システム、質問応答システムなど、自動化処理により自然言語に関するさまざまなアプリケーション/サービスが実現される。かかる自然言語処理は一般に、形態素解析、構文解析、意味解析、文脈解析という各処理フェーズに区分される。

【0099】

形態素解析では、文を意味的最小単位である形態素(m o r p h e m e)に分節して品詞の認定処理を行う。構文解析では、文法規則などを基に句構造などの文の構造を解析する。文法規則が木構造であることから、構文解析結果は一般に個々の形態素が係り受け関係などを基にして接合された木構造となる。意味解析では、文中の語の語義(概念)や、語と語の間の意味関係などに基づいて、文が伝える意味を表現する意味構造を求めて、意味構造を合成する。また、文脈解析では、文の系列である文章(談話)を解析の基本単位とみなして、文間の意味的なまとまりを得て談話構造を構成する。

【0100】

構文解析及び意味解析は、自然言語処理の分野において、対話システム、機械翻訳、文書校正支援、文書要約などのアプリケーションを実現する上で必要不可欠の技術であるとされている。

【0101】

構文解析では、自然言語文を受け取り、文法規則に基づいて単語(文節)間の係り受け関係を決定する処理を行う。構文解析結果は、依存構造と呼ばれる木構造(依存木)の形態で表現することができる。また、意味解析では、単語(文節)間の係り受け関係に基づいて文中の格関係を決定する処理を行うことができる。ここで言う格関係とは、文を構成する各要素が持つ、主語(SUBJ)、目的語(OBJ)といった文法上の役割のことを指す。また、文の時制や様相、話法などを判定する処理を意味解析が含む場合もある。

【0102】

10

20

30

40

50

構文意味解析システム例については、例えば、「増市，大熊，"Lexical Functional Grammarに基づく実用的な日本語解析システムの構築"，自然言語処理，Vol. 10，No. 2，pp. 79 - 109 (2003)」にLFGに基づく自然言語処理システムの詳細が記述されている。本発明の質問応答システムの質問意味分析手段203における質問文のQF(野球選手)と、このQFを修飾する部分「広島」の検出には、例えばこのLFGに基づく自然言語処理システムを適用することが可能である。

【0103】

図5に、Lexical Functional Grammar (LFG) LFGに基づく自然言語処理を実行する構文意味解析システム300の構成を示す。形態素解析部302は、日本語など特定の言語に関する形態素ルール302Aと形態素辞書302Bを持ち、入力文を意味的最小単位である形態素に分節して品詞の認定処理を行う。例えば、「私の娘は英語を話します。」という文が入力された場合、形態素解析結果として、「私 {Noun} の {up} 娘 {Noun} は {up} 英語 {Noun} を {up} 話す {Verb1} {tr} ます {j p} 。 {p t}」が出力される。

【0104】

このような形態素解析結果は、次いで、構文・意味解析部303に入力される。構文・意味解析部303は、文法ルール303Aや結合価辞書303Bなどの辞書を持ち、文法ルールなどに基づく句構造の解析や、文中の語の語義や語と語の間の意味関係などに基づいて文が伝える意味を表現する意味構造の解析を行う(結合価辞書は動詞と主語などの文中の他の構成要素との関係を記述したものであり、述部とそれに係る語の意味関係を抽出することができる)。そして、構文解析した結果として、単語や形態素などからなる文章の句構造を木構造として表した"c - structure (constituent structure)"と、主語、目的語などの格構造に基づいて入力文を疑問文、過去形、丁寧文など意味的・機能的に解析した結果として"f - structure (functional structure)"を出力する。

【0105】

すなわち、c - structureは、自然言語文の構造を、文の形態素を上位のフレーズへとまとめあげることによって木構造として表現するものであり、f - structureは、文法機能の概念に基づき、文の格構造、時制、様相、話法などの意味情報を属性 属性値のマトリックス構造で表現するものである。

【0106】

例えば、本例では、クライアントから入力された入力質問は、

(入力質問)「2003年に広島に行った野球選手は誰ですか？」

である。この質問文に対して、回答候補を持つ分として想定される文

想定文「2003年に広島に行った野球選手はAです」

の構文意味解析結果としてのf - structureを図6に示す。f - structureは、文法的な機能を明確に表現したものであり、文法的な機能名、意味的形式、並びに特徴シンボルにより構成される。f - structureを参照することにより、主語 (subject)、目的語 (object)、補語 (complement)、修飾語 (adjunct)といった意味理解を得ることができる。f - structureは、木構造として示されるc - structureの各節点に付随する素性の集合であり、図6に示すように属性 - 属性値のマトリックスの形で表現される。すなわち、[]で囲まれた中の左側は素性(属性)の名前であり、右側は素性の値(属性値)である。

【0107】

図6に示すf - structureは、質問文に対する回答想定文の解析結果であり、この解析結果から、質問文のQF(野球選手)と、このQFを修飾する部分「広島」の検出が実行可能となる。

【0108】

なお、質問

10

20

30

40

50

「2003年に広島に行った野球選手は誰ですか？」

の構文意味構造は、要約すると以下ようになる。

- ・述語「です」
- ・主語「2003年に広島に行った野球選手」という名詞句。
- ・XCOMP（英語のTo不定詞に相当する表現）「A」であり主語に係る。

【0109】

次に、質問文の語彙間の格フレームを得る。この時、文のメインの動詞を定め、その格要素を得る。この時、以下のヒューリスティックスを用いる。構文意味構造については図6を参照。

- * 文のメインの動詞は構文意味構造上、最も上層の一般動詞とする。 10
- * 質問文を解析した結果の「です」など一般動詞でない動詞は用いない。
- * また、XCOMP文、連体修飾節が存在する場合には、被修飾名詞を、XCOMP文あるいは連体修飾節中の動詞の格要素として認識する。
- * 格要素とならない時間などの表現は、除外する。

【0110】

上記例の場合は、以下のように格フレームが定まる。この時「広島」の意味として構文意味解析結果から「organization」が得られる。ただし、この結果は格フレームを考慮していない。

（野球選手 主格）（広島 斜格）行く。

【0111】

次に、得られた格フレームの構造を元に質問文の語彙および語彙間の意味関係を取得する。この処理においては、例えば、「「Kentaro Torisawa, "A Nearly Unsupervised Learning Method for Automatic Paraphrasing of Japanese Noun Phrases", In Proceedings of the Workshop on Automatic Paraphrasing, pp. 63 - 72, Tokyo, Japan, December, 2001」の公知の技術を用いて得た格フレーム辞書に相当する確率分布や、「河原大輔, 黒橋禎夫, "用言と直前の格要素の組を単位とする格フレームの自動構築", 自然言語処理, Vol. 9, No. 1, pp. 3 - 19, 2002」による格フレーム辞書を用いてもよい。 30

【0112】

例えば、鳥澤らの格フレーム辞書に相当する確率分布を用いた手法で、以下の格フレームを表現することを考える。

（野球選手 主格）（広島 斜格）行った。

【数1】

$P(\text{野球選手,が,広島,に,行く})$

$$= \sum_{a,b} P(\langle \text{が,に,行く} \rangle | a,b) P(\text{野球選手} | a) P(\text{広島} | b) P(a,b)$$

40

【0113】

上記式の、aとbはコーパスから確率分布を推定するときに定められる意味クラスに相当する表現である。「広島」が属しやすい意味クラスとしては「岩手, 大阪, 神戸, 横浜」などの語彙が含まれる地名のクラスや、「巨人, 横浜, 阪神, 早稲田」といったスポーツチームのクラスなど複数存在する。例えば、 $P(\text{広島} | \text{地名クラス})$ や $P(\text{広島} | \text{スポーツチームクラス})$ といった確率値は $P(\text{広島} | \text{人名クラス})$ の確率値よりも高くなる。 50

【0114】

この性質を用いて、P(野球選手, が, 広島, に, 行く)の確率値が最大となるような意味クラスの組{a, b}を選択することで、「野球選手が広島に行った」の格フレームの意味的な構造を捉えることができる。このような意味クラスの設定は、粒度の細かいシソーラス辞書と格フレーム辞書を用いることで得ることも同様に可能である。例えば、河原らの得たシソーラス辞書には、図7に示すような「行く」に関するスロットがあると

【0115】

この時、「広島」や「野球選手」という単語がどちらの「行く」に属するのかを、「河原大輔, 黒橋禎夫, "用言と直前の格要素の組を単位とする格フレームの自動構築", 自然言語処理, Vol. 9, No. 1, pp. 3 - 19, 2002」による格フレーム辞書を用いて判定することができる。格フレームの構造を用いることで、以下のように、検索クエリの拡張や、回答候補の検査に用いることができる。ここで得られたaとbの組は、「選手」クラス、「スポーツチーム」のクラスであるとして説明を進める。

【0116】

(検索クエリと検査語彙の動詞に関する拡張)

P(<が, に, V> | 選手クラス, スポーツチーム)で表される確率分布において「行く」と同様の確率値を与える動詞を得て、検索クエリを拡張することができる。例えば、「入団する」などが得られる。一般的な形式の格フレーム辞書でも、このような動詞についてのクエリ拡張を、格フレーム辞書を構成する格要素動詞のシソーラス上の距離を比較することで得ることができる。例えば、「野球選手または選手」がガ格に存在し、「広島(地名や組織)」がニ格のフレームに存在するような動詞を、格フレーム辞書から検索することで、質問文の動詞についての「入団する」などの動詞が取得でき、これを用いても回答候補の検査や、文書検索を行うことができる。

【0117】

(回答候補の検査)

回答候補が得られた文脈から、回答候補を含んで動詞と斜格の格要素と主格の格要素を取得し、主格の意味クラスを「選手」、斜格の格要素を「スポーツチーム」に固定した場合の確率値がスレッシュホールドを超えている場合に、その回答候補は検査をパスするものとする。(ただし、得られた回答候補の語彙がコーパス中に一定の頻度以上で存在しない場合には、その上位クラスとして「人名」クラスや「組織」クラスを設定するなどの工夫は必要である)。例えば、回答候補「黒田」が「黒田は広島カープに入団した。」という文から得られていた場合に、P(が, に, 入団する | 名前クラス, スポーツクラス)とP(広島カープ | スポーツチームクラス)とP(黒田 | 名前)の確率値の積がスレッシュホールドを超えている場合に検査をパスするものとする。この実施例では、確率値の積による検査を行っているが、一般的な格フレーム辞書とシソーラス辞書、さらには、意味クラスと語彙との間に人手で定めたルールを用いてもよい。

【0118】

また、「広島に行ったのは誰ですか?」のように明確にQFとなる文字列が得られない場合にも、格フレームの構造から主格と斜格の意味クラスのパターンからQFに相当するものの候補を得ることができる。すなわち、「野球選手」のような明確な文字列による質問の主題は得られないが、この質問の答えとして、格フレームから、以下のような条件の回答候補を選ぶことで可能である。

【0119】

「行く、入団するといった動詞」の主格として「選手クラスに属する名詞」をとり、かつ、斜格に「スポーツチームクラスをとる名詞」をとるなどの情報を得て、QFが取得できなくても検査が可能となる。

【0120】

このように、格フレームを制約として用いることで、格フレームの格要素から、動詞についてのクエリの拡張が可能となる。また、語彙間の意味のパターンについて格フレーム

10

20

30

40

50

照合することで、表記のゆれの吸収を動詞にまで拡大して検査を行うことができる。また、「広島に行ったのは誰ですか?」のように明確にQ Fとなる文字列が得られない場合にも、格フレームの構造から主格と斜格の意味クラスのパターンを得ることができ、回答候補に対する検査が可能となる。

【0121】

最後に、上述した処理を実行する質問応答システムを構成する情報処理装置のハードウェア構成例について、図8を参照して説明する。CPU (Central Processing Unit) 501は、OS (Operating System)に対応する処理や、上述の実施例において説明した入力質問に基づく特徴語抽出、検索処理、クエリ生成処理、パッセージ検索処理、形態素解析処理、回答候補検査処理などを実行する。これらの処理は、各情報処理装置のROM、ハードディスクなどのデータ記憶部に格納されたコンピュータ・プログラムに従って実行される。

10

【0122】

ROM (Read Only Memory) 502は、CPU 501が使用するプログラムや演算パラメータ等を格納する。RAM (Random Access Memory) 503は、CPU 501の実行において使用するプログラムや、その実行において適宜変化するパラメータ等を格納する。これらはCPUバスなどから構成されるホストバス504により相互に接続されている。

【0123】

ホストバス504は、ブリッジ505を介して、PCI (Peripheral Component Interconnect/Interface)バスなどの外部バス506に接続されている。

20

【0124】

キーボード508、ポインティングデバイス509は、ユーザにより操作される入力デバイスである。ディスプレイ510は、液晶表示装置またはCRT (Cathode Ray Tube) などから成り、各種情報をテキストやイメージで表示する。

【0125】

HDD (Hard Disk Drive) 511は、ハードディスクを内蔵し、ハードディスクを駆動し、CPU 501によって実行するプログラムや情報を記録または再生させる。ハードディスクは、例えば検索結果としての回答候補、パッセージの格納、さらに、回答候補検査手段において適用するクエリの格納、クエリーに基づくヒット文の格納、ヒット文に対する形態素解析結果の格納、回答候補の格納手段などに利用され、さらに、データ処理プログラム等、各種コンピュータ・プログラムが格納される。

30

【0126】

ドライブ512は、装着されている磁気ディスク、光ディスク、光磁気ディスク、または半導体メモリ等のリムーバブル記録媒体521に記録されているデータまたはプログラムを読み出して、そのデータまたはプログラムを、インタフェース507、外部バス506、ブリッジ505、およびホストバス504を介して接続されているRAM 503に供給する。

【0127】

接続ポート514は、外部接続機器522を接続するポートであり、USB, IEEE 1394等の接続部を持つ。接続ポート514は、インタフェース507、および外部バス506、ブリッジ505、ホストバス504等を介してCPU 501等に接続されている。通信部515は、ネットワークに接続され、クライアントやネットワーク接続サーバとの通信を実行する。

40

【0128】

なお、図8に示す質問応答システムとして適用される情報処理装置のハードウェア構成例は、PCを適用して構成した装置の一例であり、本発明の質問応答システムは、図8に示す構成に限らず、上述した実施例において説明した処理を実行可能な構成であればよい。

【0129】

以上、特定の実施例を参照しながら、本発明について詳解してきた。しかしながら、本

50

発明の要旨を逸脱しない範囲で当業者が該実施例の修正や代用を成し得ることは自明である。すなわち、例示という形態で本発明を開示してきたのであり、限定的に解釈されるべきではない。本発明の要旨を判断するためには、特許請求の範囲の欄を参酌すべきである。

【0130】

なお、明細書中において説明した一連の処理はハードウェア、またはソフトウェア、あるいは両者の複合構成によって実行することが可能である。ソフトウェアによる処理を実行する場合は、処理シーケンスを記録したプログラムを、専用のハードウェアに組み込まれたコンピュータ内のメモリにインストールして実行させるか、あるいは、各種処理が実行可能な汎用コンピュータにプログラムをインストールして実行させることが可能である。

10

【0131】

例えば、プログラムは記録媒体としてのハードディスクやROM (Read Only Memory) に予め記録しておくことができる。あるいは、プログラムはフレキシブルディスク、CD-ROM (Compact Disc Read Only Memory)、MO (Magneto optical) ディスク、DVD (Digital Versatile Disc)、磁気ディスク、半導体メモリなどのリムーバブル記録媒体に、一時的あるいは永続的に格納（記録）しておくことができる。このようリムーバブル記録媒体は、いわゆるパッケージソフトウェアとして提供することができる。

【0132】

なお、プログラムは、上述したようなリムーバブル記録媒体からコンピュータにインストールする他、ダウンロードサイトから、コンピュータに無線転送したり、LAN (Local Area Network)、インターネットといったネットワークを介して、コンピュータに有線で転送し、コンピュータでは、そのようにして転送されてくるプログラムを受信し、内蔵するハードディスク等の記録媒体にインストールすることができる。

20

【0133】

なお、明細書に記載された各種の処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的あるいは個別に実行されてもよい。また、本明細書においてシステムとは、複数の装置の論理的集合構成であり、各構成の装置が同一筐体内にあるものには限らない。

【産業上の利用可能性】

30

【0134】

以上、説明したように、本発明の構成によれば、質問に対する回答を提供するシステムにおいて、入力質問の語彙意味を解析し、入力質問の構成語の語彙意味に基づく回答候補検査処理を実行する構成としたので、質問に対する最適な回答を回答候補として選択してクライアントに提供することが可能となる。

【0135】

本発明の構成によれば、質問に対する回答を提供するシステムにおいて、入力質問に基づく検索処理結果に基づいて初期回答候補を抽出し、抽出した回答候補についての検査処理を実行し、質問に対する最適な回答を選択可能としたものであり、回答候補を適用したクエリによって検索を行い、検索結果として得られた文の構成語が、入力質問の構成語の語彙意味と同様の語彙意味を有するか否かを判定し、同様の語彙意味を有すると判定されたクエリに含まれる回答候補を、最終的な回答候補として選択する回答候補検査処理を実行する構成としたので、質問に対する最適な回答を回答候補として選択してクライアントに提供することが可能となる。

40

【図面の簡単な説明】

【0136】

【図1】本発明の質問応答システムの適用例を示すネットワーク構成図である。

【図2】本発明の一実施形態に係る質問応答システムの構成について説明する図である。

【図3】本発明の一実施形態に係る質問応答システムにおける回答候補検査手段の処理による形態素解析結果の例について説明する図である

50

【図4】本発明の一実施形態に係る質問応答システムの実行する処理シーケンスについて説明するフローチャートを示す図である。

【図5】本発明の一実施形態に係る質問応答システムにおける検査式生成手段を構成する構文意味解析システムの構成例を示す図である。

【図6】構文意味解析処理によって生成される構文意味解析結果としての `f - s t r u c t u r e` のデータ例を示した図である。

【図7】本発明の一実施形態に係る質問応答システムにおいて適用するシソーラス辞書のデータ例について説明する図である。

【図8】本発明の一実施形態に係る質問応答システムのハードウェア構成例について説明する図である。

10

【符号の説明】

【0137】

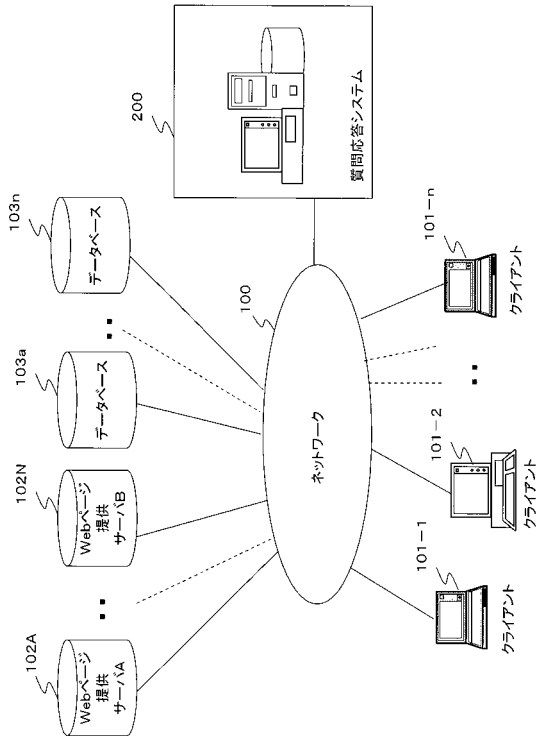
- 100 ネットワーク
- 101 クライアント
- 102 Webページ提供サーバ
- 103 データベース
- 200 質問応答システム
- 201 質問入力手段
- 202 質問タイプ判定手段
- 203 質問意味分析手段
- 204 検索手段
- 205 回答候補抽出手段
- 206 回答候補検査手段
- 207 回答選択手段
- 208 回答出力手段
- 209 ユーザ対話手段
- 300 構文意味解析システム
- 302 形態素解析部
- 302A 形態素ルール
- 302B 形態素辞書
- 303 構文意味解析部
- 303A 文法ルール
- 303B 結合価辞書
- 501 CPU (Central Processing Unit)
- 502 ROM (Read-Only-Memory)
- 503 RAM (Random Access Memory)
- 504 ホストバス
- 505 ブリッジ
- 506 外部バス
- 507 インタフェース
- 508 キーボード
- 509 ポインティングデバイス
- 510 ディスプレイ
- 511 HDD (Hard Disk Drive)
- 512 ドライブ
- 514 接続ポート
- 515 通信部
- 521 リムーバブル記録媒体
- 522 外部接続機器

20

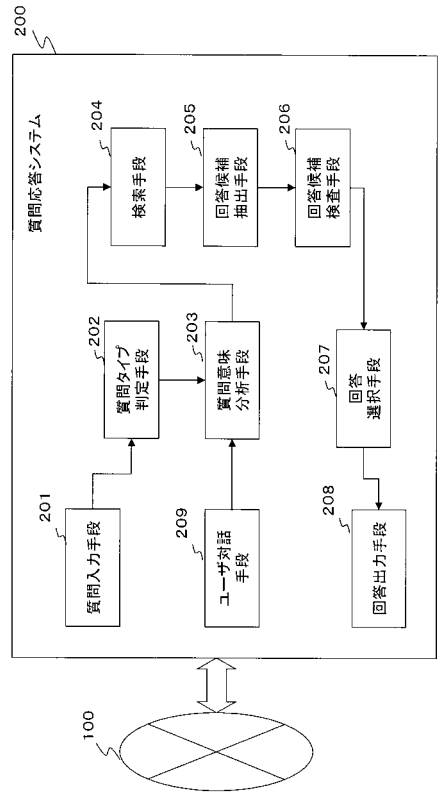
30

40

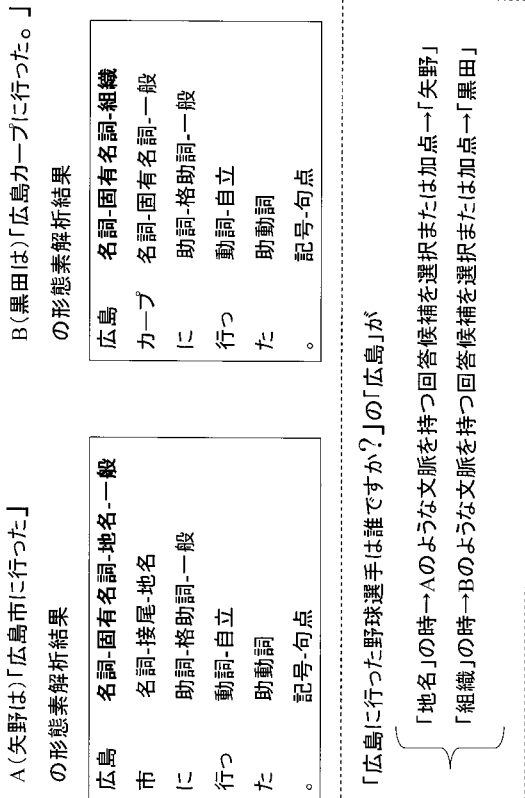
【図 1】



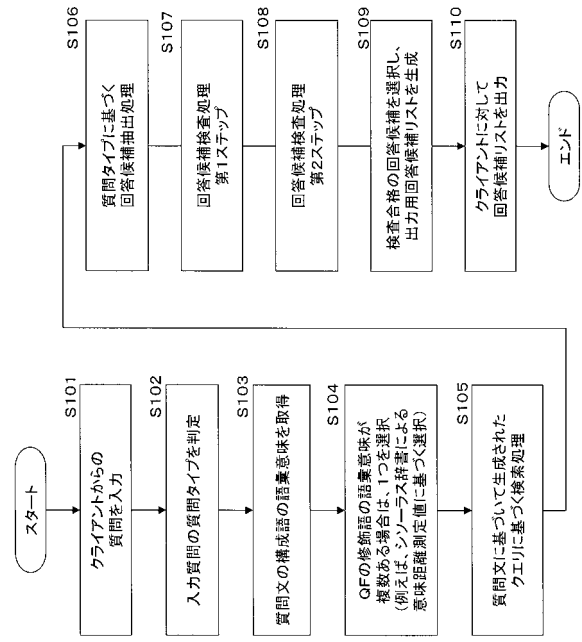
【図 2】



【図 3】



【図 4】



フロントページの続き

(72)発明者 吉村 宏樹

神奈川県足柄上郡中井町境4 3 0 グリーンテクなかい 富士ゼロックス株式会社内

(72)発明者 大熊 智子

神奈川県足柄上郡中井町境4 3 0 グリーンテクなかい 富士ゼロックス株式会社内

Fターム(参考) 5B075 KK07 ND03 ND40 NK32 NK35 NR02 NR20 PP24 PQ02 PQ12

PQ74 QM06 QP01 QP03

5B091 AA15 CA05 CA12