(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5229782号
(P5229782)
(45)発行日　平成25年7月3日(2013.7.3)　(24)登録日　平成25年3月29日(2013.3.29)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 180A
G06F 17/30 210D
G06F 17/30 330C

請求項の数 9　(全 46 頁)

(21)出願番号　特願2007-289613(P2007-289613)
(22)出願日　平成19年11月7日(2007.11.7)
(65)公開番号　特開2009-116662(P2009-116662A)
(43)公開日　平成21年5月28日(2009.5.28)
審査請求日　平成22年10月13日(2010.10.13)

特許法第３０条第１項適用　平成１９年５月１５日　国立情報学研究所発行の「ＮＴＣＩＲワークショップ　６　ミーティング」に発表　平成１９年５月１５日　ｈｔｔｐ：／／ｒｅｓｅａｒｃｈ．ｎｉｉ．ａｃ．ｊｐ／ｎｔｃｉｒ／ｗｏｒｋｓｈｏｐ／ＯｎｌｉｎｅＰｒｏｃｅｅｄｉｎｇｓ６／ＮＴＣＩＲ／１４．ｐｄｆ　を通じて発表

(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　100115749
弁理士　谷川　英和
(74)代理人　100121223
弁理士　森本　悟道
(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

審査官　打出　義尚

最終頁に続く

(54)【発明の名称】質問応答装置、質問応答方法、及びプログラム

(57)【特許請求の範囲】
【請求項１】
非ファクトイド（Ｎｏｎ－Ｆａｃｔｏｉｄ）型の質問を示す情報である質問情報を受け付ける質問情報受付部と、
前記質問情報受付部が受け付けた質問情報に対して、当該質問情報の分類を示す情報であり、理由を尋ねる質問である理由質問が少なくとも一の分類として含まれる情報である複数の分類情報のいずれかを付与する分類部と、
前記質問情報受付部が受け付けた質問情報から、用語を抽出する用語抽出部と、
分類を示す情報である分類情報と、前記用語抽出部が抽出した用語に追加する追加用語とを対応付けて有する情報である対応情報が記憶される対応情報記憶部と、
前記分類部が付与した分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出部が抽出した用語と、アクセス可能なコーパス記憶部で記憶されているコーパスと、前記分類部によって付与された分類情報に応じた式とを用いることによって、前記質問情報に対応する回答を示す情報である回答情報を前記コーパスから取得する回答情報取得部と、
前記回答情報取得部が取得した回答情報を出力する回答情報出力部と、を備え、
前記回答情報取得部が用いる式である第１の式は、前記コーパスに含まれる文書において、２個の用語が近い位置にあるほど高い値となる式であり、
前記回答情報取得部は、前記コーパスに含まれる文書について、前記分類部によって付与された分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出部が抽出

した用語とを含む用語セットから選択された２個を用いて前記式の値を算出し、当該式の値が他に比べて大きい情報である回答情報を取得し、
前記回答情報取得部は、
ある用語がある文書を特徴付けている程度を示す式である第２の式を用いて、前記用語抽出部が抽出した用語によって特徴付けられている程度の高い複数の文書を前記コーパスから取得する文書取得手段と、
前記分類部によって付与された分類情報に対応する追加用語を、前記対応情報から取得する追加用語取得手段と、
前記用語抽出部が抽出した用語と、前記追加用語取得手段が取得した追加用語とを用いて、前記文書取得手段が取得した各文書に含まれる回答情報の候補となる情報である回答候補情報について、前記分類部によって付与された分類情報に応じた前記第１の式の値を算出する算出手段と、
前記複数の回答候補情報から、前記算出手段が算出した値が他に比べて大きい値である回答情報を選択する回答情報選択手段と、を備え、
前記算出手段は、回答候補情報ｄについて、次式
【数１】

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

＋(回答候補情報が長い場合にスコアを高くするための項)

$$T3=\{t|t\in T,2dist(t1,t)\frac{df(t)}{N}\le 1\}$$

のＳｃｏｒｅ（ｄ）を算出し、当該算出値を分類情報に応じて変更した値である第１の式の値を算出する（ただし、ｄは回答候補情報であり、Ｔは、前記用語セットであり、ｄｉｓｔ（ｔ１，ｔ２）は、用語ｔ１，ｔ２の間隔であり、Ｎは、文書の総数であり、ｄｆ（ｔ）は、用語ｔの出現する文書数であり、w_{dr2}（ｔ２）は、実験によって定められる用語ｔ２の関数である）、質問応答装置。
【請求項２】
前記分類部は、前記質問情報受付部が受け付けた質問情報を、少なくとも、定義を尋ねる質問である定義質問、理由を尋ねる質問である理由質問、方法を尋ねる質問である方法質問に分類する、請求項１記載の質問応答装置。
【請求項３】
前記用語抽出部は、前記分類部によって定義質問であると分類された質問情報から、定義を尋ねている対象となる表現であるフォーカス表現の抽出も行うものであり、
前記第１の式は、
前記文書取得手段が取得した文書に前記フォーカス表現が含まれる場合には、前記フォーカス表現が含まれない場合よりも値が大きくなる式であり、
前記文書取得手段が取得した文書に含まれる前記フォーカス表現が、連体修飾節で修飾されている場合には、そうでない場合よりも値が大きくなる式であり、
前記回答情報選択手段は、前記文書取得手段が取得した文書に含まれる前記フォーカス表現が連体修飾節で修飾されている場合に、前記回答候補情報から、当該連体修飾節を回答情報として選択する、請求項２記載の質問応答装置。
【請求項４】
前記回答情報選択手段は、
前記算出手段が算出した値が他に比べて大きい値である回答候補情報を選択し、
あらかじめ用意された、質問情報と、当該質問情報の示す質問への回答を示す情報である回答情報と、当該回答情報の適否を示す情報とを少なくとも教師データとして用いて機械学習を行い、
当該機械学習の結果を用いて、前記選択した回答候補情報から回答情報を抽出する、請求

項１から請求項３のいずれか記載の質問応答装置。
【請求項５】
前記分類部は、
分類を示す情報である分類情報と、語句を示す情報である語句情報とを対応付けて有する情報である分類対応情報を記録媒体で保持しており、
前記質問情報に、語句情報が示す語句が含まれる場合に、当該質問情報に対して、当該語句情報に対応する分類情報を付与する、請求項１から請求項４のいずれか記載の質問応答装置。
【請求項６】
前記分類部は、
あらかじめ用意された、質問情報と、当該質問情報の分類を示す情報である分類情報とを教師データとして機械学習を行い、
当該機械学習の結果を用いて、前記質問情報受付部が受け付けた質問情報に対して分類情報を付与する、請求項１から請求項４のいずれか記載の質問応答装置。
【請求項７】
前記用語抽出部は、前記質問情報を形態素解析し、当該質問情報から、（１）自立語、（２）名詞、（３）名詞と動詞、（４）名詞と形容詞、（５）名詞と動詞と形容詞、から選択される（１）～（５）のいずれかに含まれる品詞の用語を抽出する、請求項１から請求項６のいずれか記載の質問応答装置。
【請求項８】
質問情報受付部と、分類部と、用語抽出部と、分類を示す情報である分類情報と、前記用語抽出部が抽出した用語に追加する追加用語とを対応付けて有する情報である対応情報が記憶される対応情報記憶部と、文書取得手段、追加用語取得手段、算出手段、及び回答情報選択手段を有する回答情報取得部と、回答情報出力部とを用いて処理される質問応答方法であって、
前記質問情報受付部が、非ファクトイド（Ｎｏｎ－Ｆａｃｔｏｉｄ）型の質問を示す情報である質問情報を受け付ける質問情報受付ステップと、
前記分類部が、前記質問情報受付ステップで受け付けた質問情報に対して、当該質問情報の分類を示す情報であり、理由を尋ねる質問である理由質問が少なくとも一の分類として含まれる情報である複数の分類情報のいずれかを付与する分類ステップと、
前記用語抽出部が、前記質問情報受付ステップで受け付けた質問情報から、用語を抽出する用語抽出ステップと、
前記回答情報取得部が、前記分類ステップで付与した分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出ステップで抽出した用語と、アクセス可能なコーパス記憶部で記憶されているコーパスと、前記分類ステップで付与された分類情報に応じた式とを用いることによって、前記質問情報に対応する回答を示す情報である回答情報を前記コーパスから取得する回答情報取得ステップと、
前記回答情報出力部が、前記回答情報取得ステップで取得した回答情報を出力する回答情報出力ステップと、を備え、
前記回答情報取得ステップで用いる式である第１の式は、前記コーパスに含まれる文書において、２個の用語が近い位置にあるほど高い値となる式であり、
前記回答情報取得ステップでは、前記コーパスに含まれる文書について、前記分類ステップにおいて付与された分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出ステップで抽出した用語とを含む用語セットから選択された２個を用いて前記式の値を算出し、当該式の値が他に比べて大きい情報である回答情報を取得し、
前記回答情報取得ステップは、
前記文書取得手段が、ある用語がある文書を特徴付けている程度を示す式である第２の式を用いて、前記用語抽出ステップで抽出した用語によって特徴付けられている程度の高い複数の文書を前記コーパスから取得する文書取得ステップと、
前記追加用語取得手段が、前記分類ステップにおいて付与された分類情報に対応する追加

用語を、前記対応情報から取得する追加用語取得ステップと、
前記算出手段が、前記用語抽出ステップで抽出した用語と、前記追加用語取得ステップで取得した追加用語とを用いて、前記文書取得ステップで取得した各文書に含まれる回答情報の候補となる情報である回答候補情報について、前記分類ステップにおいて付与された分類情報に応じた前記第１の式の値を算出する算出ステップと、
前記回答情報選択手段が、前記複数の回答候補情報から、前記算出ステップで算出した値が他に比べて大きい値である回答情報を選択する回答情報選択ステップと、を備え、
前記算出ステップでは、回答候補情報ｄについて、次式
【数２】

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

＋(回答候補情報が長い場合にスコアを高くするための項)

$$T3=\{t|t\in T,2dist(t1,t)\frac{df(t)}{N}\le 1\}$$

のＳｃｏｒｅ（ｄ）を算出し、当該算出値を分類情報に応じて変更した値である第１の式の値を算出する（ただし、ｄは回答候補情報であり、Ｔは、前記用語セットであり、ｄｉｓｔ（ｔ１，ｔ２）は、用語ｔ１，ｔ２の間隔であり、Ｎは、文書の総数であり、ｄｆ（ｔ）は、用語ｔの出現する文書数であり、w_{dr2}（ｔ２）は、実験によって定められる用語ｔ２の関数である）、質問応答方法。
【請求項９】
コンピュータを、
非ファクトイド（Ｎｏｎ－Ｆａｃｔｏｉｄ）型の質問を示す情報である質問情報を受け付ける質問情報受付部と、
前記質問情報受付部が受け付けた質問情報に対して、当該質問情報の分類を示す情報であり、理由を尋ねる質問である理由質問が少なくとも一の分類として含まれる情報である複数の分類情報のいずれかを付与する分類部と、
前記質問情報受付部が受け付けた質問情報から、用語を抽出する用語抽出部と、
前記分類部が付与した分類情報に、対応情報記憶部で記憶される、分類を示す情報である分類情報と、前記用語抽出部が抽出した用語に追加する追加用語とを対応付けて有する情報である対応情報で対応付けられている追加用語と、前記用語抽出部が抽出した用語と、アクセス可能なコーパス記憶部で記憶されているコーパスと、前記分類部によって付与された分類情報に応じた式とを用いることによって、前記質問情報に対応する回答を示す情報である回答情報を前記コーパスから取得する回答情報取得部と、
前記回答情報取得部が取得した回答情報を出力する回答情報出力部として機能させ、
前記回答情報取得部が用いる式である第１の式は、前記コーパスに含まれる文書において、２個の用語が近い位置にあるほど高い値となる式であり、
前記回答情報取得部は、前記コーパスに含まれる文書について、前記分類部によって付与された分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出部が抽出した用語とを含む用語セットから選択された２個を用いて前記式の値を算出し、当該式の値が他に比べて大きい情報である回答情報を取得し、
前記回答情報取得部は、
ある用語がある文書を特徴付けている程度を示す式である第２の式を用いて、前記用語抽出部が抽出した用語によって特徴付けられている程度の高い複数の文書を前記コーパスから取得する文書取得手段と、
前記分類部によって付与された分類情報に対応する追加用語を、前記対応情報から取得する追加用語取得手段と、
前記用語抽出部が抽出した用語と、前記追加用語取得手段が取得した追加用語とを用いて、前記文書取得手段が取得した各文書に含まれる回答情報の候補となる情報である回答候

補情報について、前記分類部によって付与された分類情報に応じた前記第１の式の値を算出する算出手段と、
前記複数の回答候補情報から、前記算出手段が算出した値が他に比べて大きい値である回答情報を選択する回答情報選択手段と、を備え、
前記算出手段は、回答候補情報ｄについて、次式
【数３】

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

＋(回答候補情報が長い場合にスコアを高くするための項)

$$T3=\{t|t\in T,2dist(t1,t)\frac{df(t)}{N}\le 1\}$$

のＳｃｏｒｅ（ｄ）を算出し、当該算出値を分類情報に応じて変更した値である第１の式の値を算出する（ただし、ｄは回答候補情報であり、Ｔは、前記用語セットであり、ｄｉｓｔ（ｔ１，ｔ２）は、用語ｔ１，ｔ２の間隔であり、Ｎは、文書の総数であり、ｄｆ（ｔ）は、用語ｔの出現する文書数であり、w_{dr2}（ｔ２）は、実験によって定められる用語ｔ２の関数である）、プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、非ファクトイド型の質問を受け付け、それに対する回答情報を出力する質問応答装置等に関する。
【背景技術】
【０００２】
従来、質問文を受け付け、その質問文に対する回答を出力する質問応答システムが開発されていた（例えば、特許文献１参照）。
【特許文献１】特許第３８６１１０５号公報
【発明の開示】
【発明が解決しようとする課題】
【０００３】
しかしながら、従来の質問応答装置は、主にＷｈａｔ型の質問に対して回答するシステムであることが多く、Ｈｏｗ，Ｗｈｙ型の質問に対して回答するシステムは、あまり開発されていなかった。また、従来のＨｏｗ，Ｗｈｙ型の質問応答システムは、性能が低く、その性能を向上させることが課題であった。
【０００４】
本発明は、上記課題を解決するためになされたものであり、Ｈｏｗ，Ｗｈｙ型の質問に対しても適切に回答することができる、高い性能を有する質問応答装置等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】
【０００５】
上記目的を達成するため、本発明による質問応答装置は、非ファクトイド（Ｎｏｎ－Ｆａｃｔｏｉｄ）型の質問を示す情報である質問情報を受け付ける質問情報受付部と、前記質問情報受付部が受け付けた質問情報に対して、当該質問情報の分類を示す情報であり、理由を尋ねる質問である理由質問が少なくとも一の分類として含まれる情報である複数の分類情報のいずれかを付与する分類部と、前記質問情報受付部が受け付けた質問情報から、用語を抽出する用語抽出部と、分類を示す情報である分類情報と、前記用語抽出部が抽出した用語に追加する追加用語とを対応付けて有する情報である対応情報が記憶される対応情報記憶部と、前記分類部が付与した分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出部が抽出した用語と、アクセス可能なコーパス記憶部で記憶され

ているコーパスと、前記分類部によって付与された分類情報に応じた式とを用いることによって、前記質問情報に対応する回答を示す情報である回答情報を前記コーパスから取得する回答情報取得部と、前記回答情報取得部が取得した回答情報を出力する回答情報出力部と、を備えたものである。

【０００６】

このような構成により、非ファクトイド型の質問情報に対しても、適切に回答情報を取得して出力することができる。さらに、追加用語を追加することによって、より高い性能が得られることになる。

【０００７】

また、本発明による質問応答装置では、前記回答情報取得部が用いる式である第１の式は、前記コーパスに含まれる文書において、２個の用語が近い位置にあるほど高い値となる式であり、前記回答情報取得部は、前記コーパスに含まれる文書について、前記分類部によって付与された分類情報に前記対応情報で対応付けられている追加用語と、前記用語抽出部が抽出した用語とから選択された２個を用いて前記式の値を算出し、当該式の値が他に比べて大きい情報である回答情報を取得してもよい。

このような構成により、第１の式を用いて回答情報を取得することにより、適切な回答情報の取得が行われることになる。

【０００８】

また、本発明による質問応答装置では、前記回答情報取得部は、ある用語がある文書を特徴付けている程度を示す式である第２の式を用いて、前記用語抽出部が抽出した用語によって特徴付けられている程度の高い複数の文書を前記コーパスから取得する文書取得手段と、前記分類部によって付与された分類情報に対応する追加用語を、前記対応情報から取得する追加用語取得手段と、前記用語抽出部が抽出した用語と、前記追加用語取得手段が取得した追加用語とを用いて、前記文書取得手段が取得した各文書に含まれる回答情報の候補となる情報である回答候補情報について、前記分類部によって付与された分類情報に応じた前記第１の式の値を算出する算出手段と、前記複数の回答候補情報から、前記算出手段が算出した値が他に比べて大きい値である回答情報を選択する回答情報選択手段と、を備えてもよい。

【０００９】

このような構成により、第２の式を用いて文書を取得し、その後に、第１の式を用いて、その文書に含まれる回答候補情報から回答情報を選択することにより、処理負荷の高い第１の式に関する計算量を減らすことができ、処理負荷を軽減することができうると共に、処理時間を短縮することができうる。

【００１０】

また、本発明による質問応答装置では、前記分類部は、前記質問情報受付部が受け付けた質問情報を、少なくとも、定義を尋ねる質問である定義質問、理由を尋ねる質問である理由質問、方法を尋ねる質問である方法質問に分類してもよい。

【００１１】

このような構成により、この分類に応じて追加用語を追加し、また、この分類に応じた式を用いることによって、適切に回答情報を取得することができるようになりうる。

【００１２】

また、本発明による質問応答装置では、前記用語抽出部は、前記分類部によって定義質問であると分類された質問情報から、定義を尋ねている対象となる表現であるフォーカス表現の抽出も行うものであり、前記第１の式は、前記文書取得手段が取得した文書に前記フォーカス表現が含まれる場合には、前記フォーカス表現が含まれない場合よりも値が大きくなる式であり、前記文書取得手段が取得した文書に含まれる前記フォーカス表現が、連体修飾節で修飾されている場合には、そうでない場合よりも値が大きくなる式であり、前記回答情報選択手段は、前記文書取得手段が取得した文書に含まれる前記フォーカス表現が連体修飾節で修飾されている場合に、前記回答候補情報から、当該連体修飾節を回答情報として選択してもよい。

このような構成により、定義質問に対して、より適切に回答情報を抽出することができるようになりうる。

【００１３】

また、本発明による質問応答装置では、前記回答情報選択手段は、前記算出手段が算出した値が他に比べて大きい値である回答候補情報を選択し、あらかじめ用意された、質問情報と、当該質問情報の示す質問への回答を示す情報である回答情報と、当該回答情報の適否を示す情報とを少なくとも教師データとして用いて機械学習を行い、当該機械学習の結果を用いて、前記選択した回答候補情報から回答情報を抽出してもよい。

このような構成により、機械学習を用いて、回答候補情報から回答情報を適切に抽出することができうる。

【００１４】

また、本発明による質問応答装置では、前記分類部は、分類を示す情報である分類情報と、語句を示す情報である語句情報とを対応付けて有する情報である分類対応情報を記録媒体で保持しており、前記質問情報に、語句情報が示す語句が含まれる場合に、当該質問情報に対して、当該語句情報に対応する分類情報を付与してもよい。

【００１５】

また、本発明による質問応答装置では、前記分類部は、あらかじめ用意された、質問情報と、当該質問情報の分類を示す情報である分類情報とを教師データとして機械学習を行い、当該機械学習の結果を用いて、前記質問情報受付部が受け付けた質問情報を分類してもよい。

【００１６】

また、本発明による質問応答装置では、前記用語抽出部は、前記質問情報を形態素解析し、当該質問情報から、（１）自立語、（２）名詞、（３）名詞と動詞、（４）名詞と形容詞、（５）名詞と動詞と形容詞、から選択される（１）～（５）のいずれかに含まれる品詞の用語を抽出してもよい。

【発明の効果】

【００１７】

本発明による質問応答装置等によれば、非ファクトイド型の質問に対しても適切に回答することができる。

【発明を実施するための最良の形態】

【００１８】

以下、本発明による質問応答装置について、実施の形態を用いて説明する。なお、以下の実施の形態において、同じ符号を付した構成要素及びステップは同一または相当するものであり、再度の説明を省略することがある。

【００１９】

（実施の形態１）

本発明の実施の形態１による質問応答装置について、図面を参照しながら説明する。本実地の形態による質問応答装置は、非ファクトイド型の質問を分類し、その分類結果に応じて回答を取得して出力するものである。

【００２０】

図１は、本実施の形態による質問応答装置１の構成を示すブロック図である。本実施の形態による質問応答装置１は、質問情報受付部１１と、分類部１２と、用語抽出部１３と、対応情報記憶部１４と、コーパス記憶部１５と、回答情報取得部１６と、回答情報出力部１７とを備える。

【００２１】

質問情報受付部１１は、非ファクトイド（Ｎｏｎ－Ｆａｃｔｏｉｄ）型の質問を示す情報である質問情報を受け付ける。この質問情報は、例えば、質問を示すテキストデータである。なお、質問情報受付部１１は、非ファクトイド型の質問以外の質問を受け付けてもよい。ここで、ファクトイド型の質問とは、名詞が回答となる質問である。例えば、「日本の首都はどこですか？」や、「人類が月に到達したのはいつですか？」等がファクトイ

ド型の質問である。それらの質問の回答は、「東京」や、「１９６９年７月１９日」のように名詞となる。非ファクトイド型の質問とは、ファクトイド型の質問と異なり、文書が回答となる質問である。例えば、「個人情報保護法に反対している人は、どうして反対しているのですか？」や、「世界遺産条約とは、どのような条約ですか？」等が非ファクトイド型の質問である。

【００２２】

質問情報受付部１１は、例えば、入力デバイス（例えば、キーボードやマウス、タッチパネルなど）から入力された質問情報を受け付けてもよく、有線もしくは無線の通信回線を介して送信された質問情報を受信してもよく、所定の記録媒体（例えば、光ディスクや磁気ディスク、半導体メモリなど）から読み出された質問情報を受け付けてもよい。例えば、ユーザの発した音声が音声認識された結果である質問情報を質問情報受付部１１が受け付けてもよい。なお、質問情報受付部１１は、受け付けを行うためのデバイス（例えば、モデムやネットワークカードなど）を含んでもよく、あるいは含まなくてもよい。また、質問情報受付部１１は、ハードウェアによって実現されてもよく、あるいは所定のデバイスを駆動するドライバ等のソフトウェアによって実現されてもよい。

【００２３】

また、質問情報受付部１１が受け付けた質問情報は、図示しない記録媒体において、一時的に記憶されていてもよい。後述する分類部１２や用語抽出部１３によって行われる処理で用いられる質問情報は、その記録媒体から読み出されたものであってもよい。その図示しない記録媒体への質問情報の蓄積は、質問情報受付部１１によって行われてもよく、あるいは、他の図示しない蓄積部によって行われてもよい。

【００２４】

分類部１２は、質問情報受付部１１が受け付けた質問情報に対して、その質問情報に応じた分類情報を付与する。分類情報は、質問情報の分類を示す情報である。分類情報には、理由を尋ねる質問である理由質問が少なくとも一の分類として含まれる。分類部１２は、質問情報受付部１１が受け付けた質問情報の分類を示す分類情報を複数の分類情報の中から特定し、その特定した分類情報を、その質問情報に対して付与する。分類部１２が、質問情報に分類情報を付与することは、例えば、付与する分類情報を所定の記録媒体に蓄積することであってもよく、付与する分類情報に対応付けて所定のフラグを設定することであってもよい。なお、ここでは、分類部１２が質問情報を分類する場合について説明したが、質問情報を分類することは、その質問情報に対応する回答を示す回答情報を分類することと等価である。例えば、ある質問情報を「理由質問」に分類することは、その質問情報に対応する回答情報を「理由回答」に分類することと等価である。したがって、分類部１２は、回答情報を分類している（回答情報に対して分類情報を付与している）、と言うこともできうる。

【００２５】

理由質問以外の分類としては、定義を尋ねる質問である定義質問、方法を尋ねる質問である方法質問、程度を尋ねる質問である程度質問、変化を尋ねる質問である変化質問、経緯を尋ねる質問である経緯質問等がある。すなわち、分類部１２は、質問情報受付部１１が受け付けた質問情報を、少なくとも、定義を尋ねる質問である定義質問、理由を尋ねる質問である理由質問、方法を尋ねる質問である方法質問、程度を尋ねる質問である程度質問、変化を尋ねる質問である変化質問、経緯を尋ねる質問である経緯質問に分類してもよい。あるいは、分類部１２は、それ以外の分類を行ってもよい。例えば、分類部１２は、質問情報受付部１１が受け付けた質問情報を、定義を尋ねる質問である定義質問、理由を尋ねる質問である理由質問、方法を尋ねる質問である方法質問に分類してもよい。

【００２６】

定義質問は、例えば、「Ｋ－１とはなんですか？」「Ｗｈａｔ　ｉｓ　Ｋ－１？」と言った質問である。

【００２７】

理由質問は、例えば、「個人情報保護法に反対している人は、どうして反対しているの

ですか？」「Ｗｈｙ ａｒｅ ｔｈｅ ｐｅｏｐｌｅ ｏｐｐｏｓｅｄ ｔｏ ｔｈｅ Ｐｒｉｖａｔｅ Ｉｎｆｏｒｍａｔｉｏｎ Ｐｒｏｔｅｃｔｉｏｎ Ｌａｗ？」と言った質問である。

【００２８】

方法質問は、例えば、「世界遺産は、どのようにして決めるのですか？」「Ｈｏｗ ｉｓ ａ Ｗｏｒｌｄ Ｈｅｒｉｔａｇｅ ｄｅｔｅｒｍｉｎｅｄ？」と言った質問である。

【００２９】

程度質問は、例えば、「チェルノブイリ原発事故の被害はどの程度でしたか？」「Ｈｏｗ ｅｘｔｅｎｓｉｖｅ ｗａｓ ｔｈｅ ｄａｍａｇｅ ｃａｕｓｅｄ ｂｙ Ｃｈｅｒｎｏｂｙｌ ｎｕｃｌｅａｒ ａｃｃｉｄｅｎｔ？」と言った質問である。

【００３０】

変化質問は、例えば、「少年法は、どう変わりましたか？」「Ｈｏｗ ｗａｓ ｔｈｅ ｊｕｖｅｎｉｌｅ ｌａｗ ｃｈａｎｇｅｄ？」と言った質問である。

【００３１】

経過質問は、例えば、「どのような経緯で琉球王国は、日本の一部になったのですか？」「Ｈｏｗ ｄｉｄ Ｒｙｕｋｙｕ ｃｏｍｅ ｔｏ ｂｅｌｏｎｇ ｔｏ Ｊａｐａｎ？」と言った質問である。

【００３２】

分類部１２は、例えば、（１）あらかじめ決められた規則に応じて分類情報を付与してもよく、あるいは、（２）機械学習を行うことによって分類情報を付与してもよい。分類を付与するこれらの方法について、以下、簡単に説明する。

【００３３】

（１）規則に応じて分類情報を付与する方法

分類部１２は、分類対応情報を図示しない記録媒体で保持しているものとする。ここで、分類対応情報は、分類を示す情報である分類情報と、語句を示す情報である語句情報とを対応付けて有する情報である。分類対応情報において、例えば、分類情報「定義質問」に、語句情報「とは何」「どんな」「どういう」「なにもの」「どのようなもの」「どういうこと」等が対応付けられていてもよい。また、例えば、分類情報「理由質問」に、語句情報「なぜ」「なにゆえ」「どうして」「何が理由で」「どんな理由で」等が対応付けられていてもよい。また、例えば、分類情報「方法質問」に、語句情報「どう」「どうすれば」「どうやって」「どのようにして」「いかにして」「いかに」「どんな方法で」等が対応付けられていてもよい。また、例えば、分類情報「程度質問」に、語句情報「どれくらい」「どれくらいの」「どの程度」等が対応付けられていてもよい。また、例えば、分類情報「変化質問」に、語句情報「何が違う」「どのように変わる」「どこが異なる」等が対応付けられていてもよい。また、例えば、分類情報「経緯質問」に、語句情報「どのような経緯」「どのようないきさつ」「どのようななりゆき」等が対応付けられていてもよい。そして、分類部１２は、質問情報受付部１１が受け付けた質問情報に、語句情報が示す語句が含まれる場合に、その質問情報に対して、その語句情報に対応する分類情報を付与する。例えば、質問情報に語句情報「どんな」が含まれる場合には、分類部１２は、その質問情報に対して、語句情報「どんな」に対応する分類情報「定義質問」を付与することができる。なお、質問情報の特定の品詞（例えば、疑問代名詞や、形容詞、副詞等）に、語句情報が示す語句が含まれる場合に、その質問情報に対して、その語句情報に対応する分類情報を付与してもよい。

【００３４】

なお、その規則は、分類対応情報以外の情報であってもよい。例えば、質問情報の先頭が「なぜ」であり、後端が「のか？」または「のですか？」である場合に、分類部１２は、その質問情報が「理由質問」であると判断してもよい。また、分類部１２が、その他の規則に応じて質問情報に分類を付与してもよいことは言うまでもない。

【００３５】

（２）機械学習によって分類情報を付与する方法

分類部１２は、あらかじめ用意された、質問情報と、その質問情報の分類を示す情報である分類情報とを教師データとして機械学習を行い、その機械学習の結果を用いて、質問情報受付部１１が受け付けた質問情報に対して分類情報を付与することができる。機械学習の際には、教師データに含まれる質問情報に対して形態素解析を行い、その形態素解析で得られた形態素を素性として用いてもよい。なお、素性として用いるのは、形態素の表層（文字列そのもの）のみであってもよく、表層と品詞であってもよく、表層と品詞と活用形であってもよい。また、形態素を素性として用いるのではなく、質問情報の所定数のキャラクタ（文字）を素性として用いてもよい。この場合に、あらゆる文字列を素性として用いてもよく、所定数の文字列（例えば、３文字連続の文字列）を、１文字ずつずらしたものをすべて素性として用いてもよく、文頭から始まる文字列だけを素性として用いてもよく、文末で終わる文字列だけを素性として用いてもよい。例えば、質問情報から連続する２文字や３文字等を取得し、それらを素性として用いてもよい。

機械学習としては、各種のアルゴリズムを用いることができる。このアルゴリズムの詳細については、［機械学習に関する説明］の欄で後述する。

【００３６】

また、形態素解析のシステムとして、日本語の場合には、例えば、奈良先端科学技術大学院大学で開発された「ＣｈａＳｅｎ（茶筌）」（ｈｔｔｐ：／／ｃｈａｓｅｎ．ｎａｉｓｔ．ｊｐ）等が知られている。また、英語の場合には、英単語に品詞を付与するソフトウェアとして、例えば、「ＴｎＴ」（ｈｔｔｐ：／／ｗｗｗ．ｃｏｌｉ．ｕｎｉ－ｓａａｒｌａｎｄ．ｄｅ／～ｔｈｏｒｓｔｅｎ／ｔｎｔ／）や「Ｂｒｉｌｌ　Ｔａｇｇｅｒ」（ｈｔｔｐ：／／ｗｗｗ．ｃｓ．ｊｈｕ．ｅｄｕ／～ｂｒｉｌｌ／）等が知られている。Ｂｒｉｌｌのものについては、例えば、次の文献を参照されたい。

【００３７】

文献：Ｅｒｉｃ　Ｂｒｉｌｌ、「Ｔｒａｎｓｆｏｒｍａｔｉｏｎ－Ｂａｓｅｄ　Ｅｒｒｏｒ－Ｄｒｉｖｅｎ　Ｌｅａｒｎｉｎｇ　ａｎｄ　Ｎａｔｕｒａｌ　Ｌａｎｇｕａｇｅ　Ｐｒｏｃｅｓｓｉｎｇ：Ａ　Ｃａｓｅ　Ｓｔｕｄｙ　ｉｎ　Ｐａｒｔ－ｏｆ－Ｓｐｅｅｃｈ　Ｔａｇｇｉｎｇ」、Ｃｏｍｐｕｔａｔｉｏｎａｌ　Ｌｉｎｇｕｉｓｔｉｃｓ，Ｖｏｌ．２１，Ｎｏ．４，ｐ．５４３－５６５、１９９５年

【００３８】

なお、ここでは、規則を用いて分類を付与する場合と、機械学習によって分類を付与する場合の２通りについて説明したが、結果として適切に質問情報を分類することができるのであれば、それ以外の方法で質問情報を分類してもよいことは言うまでもない。

【００３９】

用語抽出部１３は、質問情報受付部１１が受け付けた質問情報から、用語を抽出する。用語抽出部１３が抽出する用語は、後述する回答情報取得部１６による回答情報の取得で用いられるものである。したがって、用語抽出部１３は、質問情報を特徴付ける用語を抽出することが好適である。

【００４０】

用語抽出部１３は、質問情報を形態素解析し、その形態素解析した質問情報から、自立語である用語を抽出してもよい。また、用語抽出部１３は、形態素解析した質問情報から、名詞である用語を抽出してもよい。また、用語抽出部１３は、形態素解析した質問情報から、名詞である用語と、動詞である用語とを抽出してもよい。また、用語抽出部１３は、形態素解析した質問情報から、名詞である用語と、形容詞である用語とを抽出してもよい。また、用語抽出部１３は、形態素解析した質問情報から、名詞である用語と、動詞である用語と、形容詞である用語とを抽出してもよい。すなわち、用語抽出部１３は、質問情報を形態素解析し、その質問情報から、（１）自立語、（２）名詞、（３）名詞と動詞、（４）名詞と形容詞、（５）名詞と動詞と形容詞、から選択される（１）～（５）のいずれかに含まれる品詞の用語を抽出するものであってもよい。

【００４１】

また、用語抽出部１３は、あらかじめ図示しない記録媒体において保持されている専門用語を参照し、その専門用語と一致する用語が質問情報に含まれる場合に、その用語を抽出するようにしてもよい。その図示しない記録媒体で保持されている専門用語は、例えば、人手によって収集されたものであってもよく、技術用語辞典や、経済用語辞典、その他の専門用語の事典等から収集されたものであってもよく、あるいは、大規模なコーパスから機械的に取得されたものであってもよい。コーパスから機械的に専門用語を取得する場合には、例えば、技術文献を形態素解析することにより、単名詞等を抽出し、各単名詞等について専門用語である可能性を示すスコア付けを行い、高いスコアを付与された単名詞等を専門用語として取得してもよい。ここで、スコアを付与する方法として、造語能力に基づくスコア付け、出現頻度に基づくスコア付け等の複数の種類が知られている。また、それ以外の方法を用いてもよい。専門用語を取得する方法については、下記の複数の文献等において開示されており、従来から知られているため、その詳細な説明を省略する。また、専門用語リストを作成するツールとして、ＴｅｒｍＥｘｔｒａｃｔも公開されている（ｈｔｔｐ：／／ｇｅｎｓｅｎ．ｄｌ．ｉｔｃ．ｕ－ｔｏｋｙｏ．ａｃ．ｊｐ／ｔｅｒｍｅｘｔｒａｃｔ．ｈｔｍｌ）。

【００４２】

文献：Ｈｉｒｏｓｈｉ Ｎａｋａｇａｗａ、「Ａｕｔｏｍａｔｉｃ Ｔｅｒｍ Ｒｅｃｏｇｎｉｔｉｏｎ ｂａｓｅｄ ｏｎ Ｓｔａｔｉｓｔｉｃｓ ｏｆ Ｃｏｍｐｏｕｎｄ Ｎｏｕｎｓ」、Ｔｅｒｍｉｎｏｌｏｇｙ、Ｖｏｌ．６、Ｎｏ．２、ｐ．１９５－２１０、２０００

【００４３】

文献：大畑博一、中川裕志、「連接異なり語数による専門用語抽出」、情報処理学会研究報告、２０００－ＮＬ－１３６、ｐ．１１９－１２６

【００４４】

文献：中川裕志、森辰則、湯本紘彰、「出現頻度と連接頻度に基づく専門用語抽出」、自然言語処理、Ｖｏｌ．１０　Ｎｏ．１、ｐ．２７－４５、２００３年1月

【００４５】

また、用語抽出部１３は、質問情報に含まれる最も短いキーワードのみを用語として抽出してもよい。例えば、用語抽出部１３は、質問情報を単語列に分割して、そのそれぞれの単語（例えば、名詞や未知語の単語など）を用語として抽出してもよい。具体的には、「企業合併」が質問情報に含まれる場合には、用語「企業」「合併」がそれぞれ抽出されることになる。

【００４６】

また、用語抽出部１３は、あらゆるパターンのキーワードを用語として抽出してもよい。例えば、用語抽出部１３は、質問情報を単語列に分割して、その単語（例えば、名詞や未知語の単語など）そのものを用語として抽出すると共に、連続する単語列をも用語として抽出してもよい。具体的には、「企業合併」が質問情報に含まれる場合には、用語「企業」「合併」「企業合併」がそれぞれ抽出されることになる。この場合に、「企業の合併」から得られる用語と、「企業合併」から得られる用語とで差が生じるのは不公平であるとの観点から、正規化を行うことによって、その差を補償するようにしてもよい。

【００４７】

また、用語抽出部１３は、ラティスを利用して用語を抽出してもよい。例えば、用語抽出部１３は、前述した、あらゆるパターンのキーワードを用語として抽出する場合と同様に、あらゆる用語を特定し、その特定した各用語について、スコアを算出し、そのスコアの最大のものを用語として抽出してもよい。具体的には、「企業合併」が質問情報に含まれる場合には、用語「企業」「合併」「企業合併」がそれぞれ特定されることになる。そして、「企業」「合併」と、「企業合併」との両方について、スコアが算出され、前者の方のスコアが高い場合には、用語「企業」「合併」が抽出され、後者の方のスコアが高い場合には、用語「企業合併」が抽出される。なお、スコアを算出する式は、例えば、ＴＦ・ＩＤＦの式であってもよく、それに類似した、次の文献の式（２）であってもよく、そ

れ以外の式であってもよい。

【００４８】

文献：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」、自然言語処理（言語処理学会誌）、７巻２号，ｐ．１４１～１６０、２０００年４月

【００４９】

また、用語抽出部１３は、ｄｏｗｎ－ｗｅｉｇｈｔｉｎｇを利用して用語を抽出してもよい。例えば、用語抽出部１３は、前述した、あらゆるパターンのキーワードを用語として抽出する場合と同様に、あらゆる用語を特定し、最も短いキーワードはそのまま用語として抽出し、それよりも長いキーワードは重みが小さくなるように重み付けした上で用語として抽出する。具体的には、「企業合併」が質問情報に含まれる場合には、用語「企業」「合併」はそのまま抽出され、用語「企業合併」は重みが小さくなるように重み付けした上で抽出されることになる。

【００５０】

なお、最も短いキーワードのみを用語として抽出する方法、あらゆるパターンのキーワードを用語として抽出する方法、ラティスを利用して用語を抽出する方法、ｄｏｗｎ－ｗｅｉｇｈｔｉｎｇを利用して用語を抽出する方法については、上の文献「位置情報と分野情報を用いた情報検索」を参照されたい。

【００５１】

また、用語抽出部１３は、分類部１２によって定義質問であると分類された質問情報から、フォーカス表現の抽出をも行ってもよい。このフォーカス表現の抽出は、そのフォーカス表現が、後述する回答情報取得部１６による回答情報の取得の処理で用いられる場合にのみ行われてもよい。ここで、フォーカス表現とは、質問情報において、定義を尋ねている対象となる表現であり、フォーカス用語と呼ぶことも可能である。例えば、定義質問である質問情報が「遺伝子操作とは何ですか？」である場合には、定義を尋ねている対象となる表現は「遺伝子操作」であるため、用語抽出部１３は、フォーカス表現「遺伝子操作」を抽出する。

用語抽出部１３がフォーカス表現を抽出する方法としては、例えば、（１）手がかり句を用いる方法や、（２）機械学習を用いる方法がある。

【００５２】

（１）手がかり句を用いる方法

フォーカス表現を「Ｘ」とすると、定義質問である質問情報では、「Ｘとは」「Ｘは」「Ｘというのは」「Ｘって」などの表現が出てくることになる。したがって、用語抽出部１３は、あらかじめ図示しない記録媒体において、「とは」「は」「というのは」「って」などの手がかり句を保持しており、その手がかり句を検索キーとして定義質問である質問情報を検索し、ヒットした場合に、その手がかり句に先行する単語をフォーカス表現として抽出することができる。なお、手がかり句による検索を行う際に、手がかり句に先行する単語が名詞や未知語である場合にのみ、その単語をフォーカス表現として抽出するようにしてもよい。

【００５３】

（２）機械学習を用いる方法

用語抽出部１３は、あらかじめ用意された、フォーカス表現と、そのフォーカス表現を含む、定義質問である質問情報とを教師データとして機械学習を行い、その機械学習の結果を用いて、定義質問であると分類された質問情報から、フォーカス表現を抽出することができる。機械学習の際には、教師データに含まれる質問情報や、フォーカス表現に対して形態素解析を行い、その形態素解析で得られた形態素を素性として用いてもよい。なお、素性として用いるのは、形態素の表層（文字列そのもの）のみであってもよく、表層と品詞であってもよく、表層と品詞と活用形であってもよい。また、形態素を素性として用いるのではなく、質問情報の所定数のキャラクタ（文字）を素性として用いてもよい。この場合に、あらゆる文字列を素性として用いてもよく、所定数の文字列（例えば、３文字

連続の文字列）を、１文字ずつずらしたものをすべて素性として用いてもよく、文頭から始まる文字列だけを素性として用いてもよく、文末で終わる文字列だけを素性として用いてもよい。

機械学習としては、各種のアルゴリズムを用いることができる。このアルゴリズムの詳細については、［機械学習に関する説明］の欄で後述する。

【００５４】

なお、ここでは、手がかり句を用いてフォーカス表現を抽出する場合と、機械学習によってフォーカス表現を抽出する場合の２通りについて説明したが、結果として適切にフォーカス表現を抽出することができるのであれば、それ以外の方法でフォーカス表現を抽出してもよいことは言うまでもない。

【００５５】

また、用語抽出部１３が抽出した用語や、フォーカス表現は、図示しない記録媒体において、一時的に記憶されていてもよい。後述する回答情報取得部１６によって行われる処理で用いられる用語やフォーカス表現は、その記録媒体から読み出されたものであってもよい。その図示しない記録媒体への用語やフォーカス表現の蓄積は、用語抽出部１３によって行われてもよく、あるいは、他の図示しない蓄積部によって行われてもよい。

【００５６】

対応情報記憶部１４では、対応情報が記憶される。対応情報は、分類を示す情報である分類情報と、用語抽出部１３が抽出した用語に追加する用語である追加用語とを対応付けて有する情報である。この追加用語は、後述する回答情報取得部１６による回答情報を取得する処理において、用語抽出部１３が抽出した用語と共に用いられるものである。この追加用語を用いることによって、より適切に回答情報を取得することができるようになる。対応情報において、一の分類情報に、一または複数の追加用語が対応付けられる。この対応情報は、より高性能な回答情報の取得が行われるように、システムの設計者によって適宜、設定されるものである。対応情報の具体例については、後述する。なお、すべての分類情報に対して、追加用語が対応付けられていなくてもよい。すなわち、一部の分類情報に対しては、対応情報が対応付けられていなくてもよい。

【００５７】

対応情報記憶部１４に対応情報が記憶される過程は問わない。例えば、記録媒体を介して対応情報が対応情報記憶部１４で記憶されるようになってもよく、通信回線等を介して送信された対応情報が対応情報記憶部１４で記憶されるようになってもよく、あるいは、入力デバイスを介して入力された対応情報が対応情報記憶部１４で記憶されるようになってもよい。対応情報記憶部１４での記憶は、外部のストレージデバイス等から読み出した対応情報のＲＡＭ等における一時的な記憶でもよく、あるいは、長期的な記憶でもよい。対応情報記憶部１４は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディスクなど）によって実現されうる。

【００５８】

コーパス記憶部１５では、コーパスが記憶される。このコーパスは、大規模なものであることが好適である。このコーパスから、回答情報が取得されることになる。このコーパスは、例えば、新聞記事の情報であってもよく、百科事典等の情報であってもよく、ウェブで公開されている情報であってもよく、学術論文の情報であってもよく、特許の情報であってもよく、回答情報を取得するもととなりうる情報であれば、その内容を問わない。

【００５９】

コーパス記憶部１５にコーパスが記憶される過程は問わない。例えば、記録媒体を介してコーパスがコーパス記憶部１５で記憶されるようになってもよく、通信回線等を介して送信されたコーパスがコーパス記憶部１５で記憶されるようになってもよく、あるいは、入力デバイスを介して入力されたコーパスがコーパス記憶部１５で記憶されるようになってもよい。コーパス記憶部１５での記憶は、外部のストレージデバイス等から読み出したコーパスのＲＡＭ等における一時的な記憶でもよく、あるいは、長期的な記憶でもよい。コーパス記憶部１５は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディ

スクなど）によって実現されうる。

【００６０】

なお、本実施の形態では、質問応答装置１がコーパス記憶部１５を備える場合について説明するが、質問応答装置１は、コーパス記憶部１５を備えていなくてもよい。質問応答装置１がコーパス記憶部１５を備えていない場合であっても、質問応答装置１は、外部に存在するコーパスにアクセス可能であるものとする。質問応答装置１がアクセス可能な、外部に存在するコーパスは、一箇所に存在してもよく、分散して存在してもよい。例えば、コーパスがウェブで公開されている情報である場合には、後者となりうる。

【００６１】

回答情報取得部１６は、分類部１２が付与した分類情報に、対応情報記憶部１４で記憶されている対応情報で対応付けられている追加用語と、用語抽出部１３が抽出した用語と、アクセス可能なコーパス記憶部１５で記憶されているコーパスと、分類部１２によって付与された分類情報に応じた式を用いることによって、質問情報に対応する回答を示す情報である回答情報をコーパスから取得する。この回答情報取得部１６による回答情報の取得の方法については、図２で示される、回答情報取得部１６の詳細な構成を参照して後述する。なお、回答情報取得部１６が取得する回答情報の形式は問わない。回答情報は、例えば、テキストデータであってもよく、ＨＴＭＬやＸＭＬ等のマークアップ言語で記述されたデータであってもよい。また、文書の取得では、コーパスの全体を用いてもよく、あるいは、コーパスの一部を用いてもよい。後者の場合に、例えば、コーパスが特許の情報であれば、その要約のみを用いてもよい。

【００６２】

回答情報出力部１７は、回答情報取得部１６が取得した回答情報を出力する。回答情報出力部１７が回答情報を出力することによって、質問応答装置１のユーザは、質問情報に対応する回答を知ることができる。ここで、この出力は、例えば、表示デバイス（例えば、ＣＲＴや液晶ディスプレイなど）への表示でもよく、所定の機器への通信回線を介した送信でもよく、プリンタによる印刷でもよく、スピーカによる音声出力でもよく、記録媒体への蓄積でもよく、他の構成要素への引き渡しでもよい。なお、回答情報出力部１７は、出力を行うデバイス（例えば、表示デバイスやプリンタなど）を含んでもよく、あるいは含まなくてもよい。また、回答情報出力部１７は、ハードウェアによって実現されてもよく、あるいは、それらのデバイスを駆動するドライバ等のソフトウェアによって実現されてもよい。

【００６３】

図２は、回答情報取得部１６の構成を示すブロック図である。図２において、本実施の形態による回答情報取得部１６は、文書取得手段２１と、追加用語取得手段２２と、算出手段２３と、回答情報選択手段２４とを備える。

【００６４】

文書取得手段２１は、第２の式を用いて、用語抽出部１３が抽出した用語によって特徴付けられている程度の高い複数の文書をコーパスから取得する。ここで、第２の式とは、ある用語がある文書を特徴付けている程度を示す式である。例えば、第２の式は、一の文書に出現する用語の頻度が高ければ大きい値となり、かつ、多くの文書にその用語が出現するのであれば小さい値となる式であってもよい。文書取得手段２１は、コーパスと、抽出された用語とを用いて出現頻度等の値を算出し、その値を用いて第２の式の値を算出してもよい。文書取得手段２１は、その第２の式において、用語抽出部１３が抽出した用語について和をとることによって、文書ごとの第２の式の値を算出し、その値の高い文書を取得してもよい。この第２の式は、例えば、ＴＦ・ＩＤＦの式であってもよく、それを改良した次の式であってもよい。本実施の形態では、第２の式が、次の式（１）である場合について説明する。

【００６５】

【数１】

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

【００６６】

ここで、ｄは文書である。文書は、ひとまとまりの文の集合である。文書は、例えば、コーパスが新聞記事の情報である場合に、１個の記事であってもよく、コーパスが百科事典等の情報である場合に、１個の用語に関する解説であってもよく、コーパスがウェブで公開されている情報である場合に、１個のページであってもよく、コーパスが特許の情報である場合に、１個の公報であってもよく、あるいは、それらの情報に含まれる１または複数のパラグラフであってもよい。ｔは、質問情報から用語抽出部１３によって抽出された用語である。ｔｆ（ｄ，ｔ）は、文書ｄにおけるｔの出現頻度（出現回数）である。ｄｆ（ｔ）は、ｔの出現する文書数である。Ｎは、文書の総数である。ｌｅｎｇｔｈ（ｄ）は、ｄの長さである。ｄの長さは、例えば、文書ｄのバイト数や、文字数、単語数等によって示される。Δは、全文書の長さの平均である。k_t、k_+は、定数であって、実験結果によって定められる値である。例えば、k_t、k_+として、それぞれ０．０００００１，２０を用いてもよい。上記（１）式は、ロバートソンのＯｋａｐｉウェイティングの式に基づくものである。

【００６７】

また、ＴＦ・ＩＤＦの式は、次のようになる。

Ｓｃｏｒｅ（ｄ）＝Σ（ｔｆ（ｄ，ｔ）＊ｌｏｇ（Ｎ／ｄｆ（ｔ））

ただし、Σは、ｔに関する和である。また、ｔｆ（ｄ，ｔ）、ｄｆ（ｔ）等は、前述の説明と同様である。

【００６８】

なお、文書取得手段２１がＴＦ・ＩＤＦの式を用いる場合にも、上記の式（１）と同様に、用語ｔについて、和をとるものとする。また、本実施の形態では、和をとる用語ｔは、用語抽出部１３が抽出した用語である場合について説明するが、文書取得手段２１は、用語抽出部１３が抽出した用語と、分類部が付与した分類情報に対応情報で対応付けられている追加用語とを、和をとる用語（上記の式（１）や、ＴＦ・ＩＤＦの式におけるｔ）として用いてもよい。

【００６９】

文書取得手段２１は、コーパス記憶部１５に含まれる各文書ｄに対して、第２の式である上記の式（１）の値を算出する。そして、Ｓｃｏｒｅ（ｄ）の値が大きい文書ｄを取得する。Ｓｃｏｒｅ（ｄ）の値が大きい文書ｄとは、例えば、Ｓｃｏｒｅ（ｄ）の値が、しきい値よりも大きい値である文書ｄであってもよく、Ｓｃｏｒｅ（ｄ）の値が大きい方から選択された、あらかじめ決められた個数の文書ｄや、あらかじめ決められた割合の文書ｄであってもよい。しきい値よりも大きい値とは、しきい値を含んでもよく、あるいは、含まなくてもよい。また、しきい値は、例えば、あらかじめ設定された値であってもよく、算出されたＳｃｏｒｅ（ｄ）に応じて定められてもよい。後者の場合には、例えば、しきい値は、Ｓｃｏｒｅ（ｄ）の最大値に０．９を掛けた値であってもよい。

【００７０】

また、文書取得手段２１が取得した文書は、図示しない記録媒体において、一時的に記憶されていてもよい。後述する算出手段２３によって行われる処理で用いられる文書は、その記録媒体から読み出されたものであってもよい。その図示しない記録媒体への文書の蓄積は、文書取得手段２１によって行われてもよく、あるいは、他の図示しない蓄積部によって行われてもよい。また、文書取得手段２１による文書の取得は、例えば、文書を識別する情報である文書識別情報を取得することであってもよい。

【００７１】
追加用語取得手段２２は、分類部１２によって付与された分類情報に対応する追加用語を、対応情報から取得する。追加用語取得手段２２は、例えば、分類部１２による分類結果を示す分類情報を取得し、その分類情報を検索キーとして、対応情報記憶部１４において記憶されている対応情報を検索し、検索された分類情報に対応付けられている追加用語を対応情報記憶部１４から取得することによって、追加用語を取得することができる。
【００７２】
その取得された追加用語は、図示しない記録媒体において、一時的に記憶されてもよい。後述する算出手段２３によって行われる処理で用いられる追加用語は、その記録媒体から読み出されたものであってもよい。その図示しない記録媒体への追加用語の蓄積は、追加用語取得手段２２によって行われてもよく、あるいは、他の図示しない蓄積部によって行われてもよい。
【００７３】
算出手段２３は、用語抽出部１３が抽出した用語と、追加用語取得手段２２が取得した追加用語とを用いて、文書取得手段２１が取得した各文書に含まれる回答情報の候補となる情報である回答候補情報について、分類部１２によって付与された分類情報に応じた第１の式の値を算出する。ここで、第１の式とは、コーパスに含まれる文書において、２個の用語が近い位置にあるほど高い値となる式である。算出手段２３は、例えば、抽出された用語や、追加用語、回答候補情報、コーパス等を用いて、用語の近さや用語の出現する文書数等の値を算出し、その値を用いて、第１の式の値を算出してもよい。算出手段２３は、その第１の式において、用語抽出部１３が抽出した用語や、追加用語について和をとることによって、回答候補情報ごとの第１の式の値を算出してもよい。具体的には、第１の式は、次の式（２）を用いた式であってもよい。
【００７４】
【数２】

$$Score(d)=\max_{t1\in T}\sum_{t2\in T3}w_{dr2}(t2)\log\frac{N}{2dist(t1,t2)*df(t2)}+0.00000001\times length(d) \qquad (2)$$

$$T3=\{t\,|\,t\in T,2dist(t1,t)\frac{df(t)}{N}\le 1\} \qquad (3)$$

【００７５】
ここで、ｄは、回答候補情報である。回答候補情報は、文書取得手段２１が取得した文書に含まれる、あらかじめ決められた分量のテキスト情報である。回答候補情報は、例えば、文書取得手段２１が取得した文書に含まれる１パラグラフであってもよく、連続する２パラグラフであってもよく、連続する３パラグラフであってもよく、１パラグラフと、連続する２パラグラフと、連続する３パラグラフの集合であってもよく、連続する文のあらゆる組合せ（例えば、文書に含まれる第１文のみ、第１文から第２文まで、第１文から第３文まで、第１文から第４文まで、・・・、第２文のみ、第２文から第３文まで、第２文から第４文まで、第２文から第５文まで、・・・のそれぞれを回答候補情報とする場合など）であってもよく、連続する文節のあらゆる組合せ（例えば、文書に含まれる第１節から第２節まで、第１節から第３節まで、第１節から、第４節まで、・・・、第２節から第３節まで、第２節から第４節まで、第２節から第５節まで、・・・のそれぞれを回答候補情報とする場合など。その回答候補情報では、文をまたがってもよい。）であってもよく、その他の分量のテキスト情報であってもよい。Ｔは、質問情報から抽出された用語と、追加用語とを含む用語セットである。w_{dr2}（ｔ２）は、例えば、ｔ２が動詞であれば「０．５」となり、それ以外の品詞の用語であれば「１」となる。なお、w_{dr2}（ｔ２）は、それ以外の複雑な設定であってもよい。ｄｉｓｔ（ｔ１，ｔ２）は、ｔ１とｔ２との間隔である。なお、その間隔は、例えば、ｔ１とｔ２との間の文字数であってもよく

、その間に含まれる文字のバイト数であってもよく、その間に含まれる単語数であってもよい。また、便宜上、ｔ１＝ｔ２である場合に、ｄｉｓｔ（ｔ１，ｔ２）＝０．５であるとする。ｌｅｎｇｔｈ（ｄ）は、ｄの長さである。ｄの長さは、例えば、文書ｄのバイト数や、文字数、単語数等によって示される。上記の式（２）における第２項は、回答候補情報が長い場合にスコアを高くするために用いられる。また、上記の式（２）で用いられるＴ３は、式（３）を満たすｔのセットである。

【００７６】

なお、２以上の追加用語が取得された場合に、その２以上の追加用語をすべてＴに追加してもよく、あるいは、取得された２以上の追加用語のうち、１個をＴに追加して、上記の式（２）の値を算出し、次に、その追加した追加用語を削除して、取得された２以上の追加用語のうちの新たな追加用語をＴに追加して、上記の式（２）の値を算出することを繰り返して実行するようにしてもよい。さらに、同様のことを、取得された２以上の追加用語の２以上のすべての組合せに対して行ってもよい。

【００７７】

また、算出手段２３は、分類部１２による質問情報の分類結果に応じた第１の式の値を算出する。算出手段２３は、例えば、質問情報が程度質問に分類された場合であって、回答候補情報に数表現が含まれる場合に、そうでない場合よりも値の大きくなる式を用いて、第１の式の値を算出してもよい。より具体的には、算出手段２３は、質問情報が程度質問に分類された場合であって、回答候補情報に数表現が含まれる場合に、上記の式（２）の結果を１．１倍したものを、第１の式の値として算出してもよい。ここで、算出手段２３は、回答候補情報に数表現が存在するかどうかを、例えば、回答候補情報に数字が存在するかどうかによって判断してもよい。その数字は、一般にアラビア数字であるが、算出手段２３は、漢数字や、ローマ数字が含まれるかどうかについても判断し、回答候補情報に漢数字や、ローマ数字が含まれる場合にも、数表現が存在すると判断してもよい。

【００７８】

また、算出手段２３は、例えば、文書取得手段２１が取得した文書にフォーカス表現が含まれる場合（すなわち、用語抽出部１３によって質問情報からフォーカス表現が抽出されており、かつ、その抽出されたフォーカス表現が、文書取得手段２１によって取得された文書に含まれる場合）には、フォーカス表現が含まれない場合よりも値が大きくなる式を用いて、第１の式の値を算出してもよい。より具体的には、算出手段２３は、文書取得手段２１が取得した文書にフォーカス表現が含まれる場合に、上記の式（２）の結果を１．１倍したものを、第１の式の値をして算出してもよい。また、算出手段２３は、文書取得手段２１が取得した文書に含まれるフォーカス表現が、連体修飾節で修飾されている場合（すなわち、用語抽出部１３によって質問情報からフォーカス表現が抽出されており、かつ、その抽出されたフォーカス表現が、文書取得手段２１によって取得された文書に含まれており、なおかつ、文書取得手段２１によって取得された文書に含まれるフォーカス表現が、連体修飾語で修飾されている場合）には、そうでない場合よりも値が大きくなる式を用いて、第１の式の値を算出してもよい。より具体的には、算出手段２３は、文書取得手段２１が取得した文書に含まれるフォーカス表現が、連体修飾節で修飾されている場合に、上記の式（２）の結果を１．１倍したものを、第１の式の値として算出してもよい。ここで、算出手段２３は、文書取得手段２１が取得した文書にフォーカス表現が含まれる場合に、その文書に対して係り受け解析を行い、その解析結果を用いて、フォーカス表現に係っている文節を連体修飾節として特定することができる。したがって、そのフォーカス表現に係っている文節である連体修飾節が存在するのであれば、算出手段２３は、上記の式（２）の結果を１．１倍したものを、第１の式の値として算出してもよい。その係り受け解析を行うシステムとして、例えば、奈良先端科学技術大学院大学で開発された「Ｃａｂｏｃｈａ」（ｈｔｔｐ：／／ｃｈａｓｅｎ．ｏｒｇ／～ｔａｋｕ／ｓｏｆｔｗａｒｅ／ｃａｂｏｃｈａ／）等が知られている。その係り受け解析を行うシステムを用いることによって、文節の認識と、文節間の係り受けの関係を示す情報の取得とを行うことができる。なお、ここでは、付与された分類情報に応じた式として、上記の式（２）に所定の

値を乗算する場合について説明したが、付与された分類情報に応じた式は、付与された分類情報ごとにまったく異なる形式の式であってもよい。また、付与された分類情報に応じた式は、あらかじめその分類情報ごとに、図示しない記録媒体において保持されていてもよく、あるいは、図示しない記録媒体においては、基本となる式（例えば、上記の式（２））が保持されており、算出手段２３が、付与された分類情報に応じて、その基本となる式を変形させて（例えば、分類結果に応じて、その基本となる式を１．１倍するなどの変形）用いてもよい。

【００７９】

なお、算出手段２３の算出した値は、図示しない記録媒体において一時的に記憶されていてもよい。後述する回答情報選択手段２４によって行われる処理で用いられる値は、その記録媒体から読み出されたものであってもよい。その図示しない記録媒体への値の蓄積は、算出手段２３によって行われてもよく、あるいは、他の図示しない蓄積部によって行われてもよい。

【００８０】

回答情報選択手段２４は、複数の回答候補情報から、算出手段２３が算出した値が他に比べて大きい値である回答候補情報を回答情報として選択する。算出手段２３が算出した値が他に比べて大きい値である回答情報とは、例えば、算出手段２３によって算出された値が、しきい値よりも大きい値である回答候補情報であってもよく、算出手段２３によって算出された値が大きい方から選択された、あらかじめ決められた個数や、あらかじめ決められた割合の回答候補情報であってもよい。しきい値よりも大きい値とは、しきい値を含んでもよく、あるいは、含まなくてもよい。また、しきい値は、例えば、あらかじめ設定された値であってもよく、算出手段２３が算出した値に応じて定められてもよい。後者の場合には、例えば、しきい値は、算出手段２３が算出した値の最大値に０．９を掛けた値であってもよい。

【００８１】

なお、複数の回答候補情報から回答情報を選択する際に、機械学習を行い、その機械学習の結果を用いて、その選択を行ってもよい。例えば、回答情報選択手段２４は、複数の回答候補情報から、算出手段２３が算出した値が他に比べて大きい値である回答候補情報を選択し、その選択した回答候補情報から、機械学習の結果を用いて回答候補情報を選択して、その選択した回答候補情報を回答情報としてもよい。回答候補情報は、前述のように、１パラグラフや、連続する複数のパラグラフであってもよく、連続する文のあらゆる組合せであってもよく、連続する文節のあらゆる組合せであってもよく、その他の分量のテキスト情報であってもよい。その機械学習では、あらかじめ用意された、質問情報と、その質問情報の示す質問への回答を示す回答情報と、その回答情報の適否を示す情報とを少なくとも教師データとして用いて機械学習を行ってもよく、さらに、その教師データに含まれる質問情報に応じた分類情報をも、教師データとして用いて機械学習を行ってもよい。また、その回答候補情報に対応する前述の第１の式の値も教師データに含めてもよい。また、回答候補情報を含む文書のうち、回答候補情報以外の部分に含まれる文や単語、文字列等を教師データに含めてもよい。なお、機械学習の結果を用いた回答候補情報の選択の際には、教師データで用いた情報のうち、回答情報の適否を示す情報以外の情報を用いて、回答候補情報の選択を行うことになる。例えば、教師データに質問情報と、その質問情報の分類を示す分類情報と、回答情報と、その回答の適否を示す情報とが含まれる場合には、回答候補情報の選択の際にも、質問情報と、その質問情報に付与された分類情報と、回答候補情報（教師データにおける回答情報に対応する情報である）と、機械学習の結果とを用いて、その回答候補情報が回答情報として適切であるのか、あるいは、適切でないのかを判断することになる。また、機械学習の結果を用いた回答候補情報の選択において、回答候補情報が回答情報である確からしさ（確信度）の値が他に比べて大きい値である回答候補情報を回答情報として選択してもよい。確からしさの値が他に比べて大きい値である回答候補情報とは、前述の算出手段２３が算出した値が他に比べて大きい値である回答候補情報の場合と同様である。機械学習の際には、教師データに含まれる質問情報

や回答候補情報に対して形態素解析を行い、その形態素解析で得られた形態素を素性として用いてもよい。なお、素性として用いるのは、形態素の表層（文字列そのもの）のみであってもよく、表層と品詞であってもよく、表層と品詞と活用形であってもよい。また、形態素を素性として用いるのではなく、質問情報等の所定数のキャラクタ（文字）を素性として用いてもよい。この場合に、あらゆる文字列を素性として用いてもよく、所定数の文字列（例えば、３文字連続の文字列）を、１文字ずつずらしたものをすべて素性として用いてもよく、文頭から始まる文字列だけを素性として用いてもよく、文末で終わる文字列だけを素性として用いてもよい。機械学習としては、各種のアルゴリズムを用いることができる。このアルゴリズムの詳細については、［機械学習に関する説明］の欄で後述する。

【００８２】

また、回答情報選択手段２４は、文書取得手段２１が取得した文書に含まれるフォーカス表現が連体修飾節で修飾されている場合に、回答候補情報から、その連体修飾節を回答情報として選択してもよい。連体修飾節を選択する方法は、回答情報選択手段２４は、例えば、文書取得手段２１が取得した文書に対して係り受け解析を行い、その解析結果を用いて、フォーカス表現に係っている文節を連体修飾節として特定することができる。前述のように、その係り受け解析を行うシステムとして、例えば、奈良先端科学技術大学院大学で開発された「Ｃａｂｏｃｈａ」（ｈｔｔｐ：／／ｃｈａｓｅｎ．ｏｒｇ／～ｔａｋｕ／ｓｏｆｔｗａｒｅ／ｃａｂｏｃｈａ／）等が知られている。

【００８３】

回答情報選択手段２４が回答情報を選択するとは、例えば、選択した回答情報そのものを図示しない記録媒体に蓄積することであってもよく、選択した回答情報を特定する情報（例えば、回答情報の格納されているコーパスの位置を示すポインタなど）を図示しない記録媒体に蓄積することであってもよく、あるいは、選択した回答情報に対応付けてフラグ等を設定することであってもよい。

【００８４】

また、回答情報選択手段２４は、複数の回答候補情報から、算出手段２３が算出した値が他に比べて大きい値である回答候補情報を選択すると共に、機械学習を行い、その機械学習の結果を用いて、選択した回答候補情報から回答情報を抽出してもよい。その機械学習では、あらかじめ用意された、質問情報と、その質問情報の示す質問への回答を示す情報である回答情報と、その回答情報の適否を示す情報とを少なくとも教師データとして用いて機械学習を行ってもよく、さらに、その教師データに含まれる質問情報に応じた分類情報をも、教師データとして用いて機械学習を行ってもよい。また、回答情報の抽出された回答候補情報も教師データに含めてもよく、さらに、その回答候補情報に対応する前述の第１の式の値も教師データに含めてもよい。機械学習の際には、教師データに含まれる質問情報や回答情報に対して形態素解析を行い、その形態素解析で得られた形態素を素性として用いてもよい。なお、素性として用いるのは、形態素の表層（文字列そのもの）のみであってもよく、表層と品詞であってもよく、表層と品詞と活用形であってもよい。また、形態素を素性として用いるのではなく、質問情報等の所定数のキャラクタ（文字）を素性として用いてもよい。この場合に、あらゆる文字列を素性として用いてもよく、所定数の文字列（例えば、３文字連続の文字列）を、１文字ずつずらしたものをすべて素性として用いてもよく、文頭から始まる文字列だけを素性として用いてもよく、文末で終わる文字列だけを素性として用いてもよい。機械学習としては、各種のアルゴリズムを用いることができる。このアルゴリズムの詳細については、［機械学習に関する説明］の欄で後述する。

【００８５】

また、回答情報選択手段２４は、複数の回答候補情報から、算出手段２３が算出した値が他に比べて大きい値である回答候補情報を選択すると共に、選択した回答候補情報から、所定の規則に応じて、回答情報を抽出してもよい。所定の規則とは、例えば、用語抽出部１３によって抽出された用語や、追加用語取得手段２２によって取得された追加用語を

最も多く含む文やパラグラフを回答情報として抽出するとの規則であってもよく、質問情報に対して付与された分類情報に合致した表現を含む文やパラグラフを回答情報として抽出するとの規則であってもよい。後者の場合には、例えば、図示しない記録媒体において、分類情報と、回答情報に含まれているべき表現とを対応付ける情報である分類・回答対応情報が記憶されており、その情報を用いて、回答情報を抽出してもよい。より具体的には、分類・回答対応情報において、分類情報「理由質問」に、回答情報に含まれているべき表現「だからです」「だからである」等が対応付けられており、回答情報選択手段２４は、分類情報「理由質問」が付与された質問情報に対応する回答情報を選択する際には、その分類・回答対応情報を参照し、分類情報「理由質問」に対応するいずれかの表現を含む文やパラグラフを、回答情報として選択してもよい。

【００８６】

なお、本実施の形態では、回答情報取得部１６が図２で示される構成である場合について説明するが、回答情報取得部１６は、コーパスに含まれる文書について、分類部１２によって付与された分類情報に対応情報で対応付けられている追加用語と、用語抽出部１３が抽出した用語とから選択された２個を用いて第１の式の値を算出し、その第１の式の値が他に比べて大きい情報である回答情報を取得するものであれば、その構成を限定されるものではない。例えば、回答情報取得部１６は、文書取得手段２１を備えず、コーパス記憶部１５で記憶されているコーパスに含まれる各回答候補情報に対して、算出手段２３によって、第１の式の値を算出し、その値が他に比べて大きい回答候補情報を、回答情報として取得してもよい。

【００８７】

また、回答情報取得部１６は、文書取得手段２１が取得した文書をリランキングする図示しないリランキング手段をさらに備え、算出手段２３は、文書取得手段２１が取得した文書に代えて、リランキング手段によって上位にリランキングされた文書に含まれる回答候補情報について、第１の式の値を算出するようにしてもよい。上位にリランキングされた文書とは、例えば、リランキングの際に算出された値が、しきい値よりも大きい値の文書であってもよく、その算出された値が大きい方から所定数、あるいは所定割合の文書であってもよい。しきい値よりも大きい値とは、しきい値を含んでもよく、あるいは、含まなくてもよい。また、しきい値は、例えば、あらかじめ設定された値であってもよく、リランキング手段が算出した値に応じて定められてもよい。後者の場合には、例えば、しきい値は、リランキング手段が算出した値の最大値に０．９を掛けた値であってもよい。

【００８８】

ここで、その図示しないリランキング手段について説明する。リランキング手段は、次式を用いて、文書取得手段２１が取得した文書をリランキングする。

【数３】

$$Score(d)=\max_{t1\in T}\sum_{t2\in T3}w_{dr2}(t2)\log\frac{N}{2dist(t1,t2)*df(t2)} \qquad (4)$$

$$T3=\{t|t\in T,2dist(t1,t)\frac{df(t)}{N}\le 1\} \qquad (5)$$

【００８９】

この式において、ｄは文書取得手段２１が取得した文書である。また、その他のｄｉｓｔ（ｔ１，ｔ２）等は、式（２）、式（３）の説明と同様のものである。また、Ｔは、質問情報から抽出された用語のセットであってもよく、質問情報から抽出された用語と、追加用語とを含む用語セットであってもよい。後者の場合であって、２以上の追加用語が取得された場合に、その２以上の追加用語をすべてＴに追加してもよく、あるいは、取得された２以上の追加用語のうち、１個をＴに追加して、上記の式（４）の値を算出し、次に、その追加した追加用語を削除して、取得された２以上の追加用語のうちの新たな追加用語をＴに追加して、上記の式（４）の値を算出することを繰り返して実行するようにして

もよい。さらに、同様のことを、取得された２以上の追加用語の２以上のすべての組合せに対して行ってもよい。

【００９０】

このように、図示しないリランキング手段を備えることによって、文書取得手段２１が取得した文書をさらに絞り込むことができうる。例えば、文書取得手段２１が３００個の文書を取得し、リランキング手段によるリランキング（上記式（４）を用いたスコアの算出）によって、上位から２０個の文書を取得して、算出手段２３による算出で用いるようにしてもよい。前述の説明から明らかなように、算出手段２３での算出は、非常に負荷の大きい処理であるため、あらかじめ、その処理で用いる文書の数を絞り込んでおくことは、処理負荷を軽減し、処理時間を短縮するために有用である。

【００９１】

また、本実施の形態では、質問情報受付部１１が、非ファクトイド型の質問情報を受け付ける場合について主に説明するが、質問情報受付部１１は、非ファクトイド型以外の質問情報をも受け付け、非ファクトイド型の質問情報であるのか、あるいは、それ以外の質問情報であるのかの判断を行ったうえで、非ファクトイド型の質問情報であれば、本実施の形態による手法によって処理し、それ以外の質問情報であれば、従来例の手法によって処理するようにしてもよい。なお、非ファクトイド型以外の質問情報が、ファクトイド型の質問情報であるのか、あるいは、それ以外の質問情報であるのかについて、さらに判断してもよい。それらの判断は、例えば、規則に基づいた判断でもよく、あるいは、機械学習による判断でもよい。この判断は、図示しない質問情報判断部によって行われてもよい。以下、その質問情報判断部が行いうる、（１）規則に基づいて質問情報に関する判断を行う方法と、（２）機械学習によって質問情報に関する判断を行う方法について、簡単に説明する。

【００９２】

（１）規則に基づいて質問情報に関する判断を行う方法

図示しない質問情報判断部は、判断に用いる情報である判断情報を図示しない記録媒体で保持しているものとする。ここで、判断情報は、質問情報が非ファクトイド型の質問であるかどうかなどの判断を行うための規則を示す情報である。判断情報は、例えば、質問情報の種類を示す情報と、質問に含まれる語句を示す情報とを対応付けて有する情報であってもよい。判断情報において、例えば、種類を示す情報「非ファクトイド型」に、語句を示す情報「とは何」「どんな」「どういう」「どういった」「何もの」「なぜ」「なにゆえ」「どうして」「何が理由で」「どんな理由で」「どうすれば」「いかにして」「どうやって」「どのようにして」「どれくらい」「どの程度」「何がちがう」「どのように変わる」「どこが異なる」「どのような経緯」「どのようないきさつ」「どのようななりゆき」等が対応付けられていてもよい。また、種類を示す情報「ファクトイド型」に、語句を示す情報「いつ」「誰が」「どこ」「読みは何」等が対応付けられていてもよい。そして、質問情報判断部は、質問情報受付部１１が受け付けた質問情報に、語句を示す情報が含まれる場合に、その質問情報が、その語句を示す情報に対応する種類の質問であると判断してもよい。なお、いずれの語句も含まれない場合には、ファクトイド型の質問でもなく、非ファクトイド型の質問でもない、その他の質問であると判断してもよい。

【００９３】

（２）機械学習によって質問情報に関する判断を行う方法

図示しない質問情報判断部は、あらかじめ用意された、質問情報と、その質問情報の種類を示す情報とを教師データとして機械学習を行い、その機械学習の結果を用いて、質問情報受付部１１が受け付けた質問情報が、ファクトイド型の質問であるのか、非ファクトイド型の質問であるのか、あるいは、その他の質問であるのかを判断してもよい。機械学習の際には、教師データに含まれる質問情報に対して形態素解析を行い、その形態素解析で得られた形態素を素性として用いてもよい。なお、素性として用いるのは、形態素の表層（文字列そのもの）のみであってもよく、表層と品詞であってもよく、表層と品詞と活用形であってもよい。また、形態素を素性として用いるのではなく、質問情報の所定数の

キャラクタ（文字）を素性として用いてもよい。この場合に、あらゆる文字列を素性として用いてもよく、所定数の文字列（例えば、３文字連続の文字列）を、１文字ずつずらしたものをすべて素性として用いてもよく、文頭から始まる文字列だけを素性として用いてもよく、文末で終わる文字列だけを素性として用いてもよい。例えば、質問情報から連続する２文字や３文字等を取得し、それらを素性として用いてもよい。

機械学習としては、各種のアルゴリズムを用いることができる。このアルゴリズムの詳細については、［機械学習に関する説明］の欄で後述する。

【００９４】

ここで、図示しない質問情報判断部によって、非ファクトイド型の質問であると判断された質問情報については、前述のように、本実施の形態による質問応答装置１による処理、すなわち、分類部１２による分類や、用語抽出部１３による用語抽出の処理が行われていくことになる。一方、ファクトイド型の質問であると判断された質問情報については、従来例の方法を用いて、回答情報が取得されるようにしてもよい。また、ファクトイド型の質問情報に対応する回答情報を取得する方法として、例えば、次のような方法を用いてもよい。

【００９５】

［ファクトイド型の質問情報に対応する回答情報を抽出する方法］

まず、質問情報に対応する回答情報がどのような解表現になるのかを推定する。例えば、回答情報が「人名」になるのか、「場所」になるのか、「時間表現」になるのか、「数値表現」になるのか、「国名」になるのか、「首都名」になるのか、「平仮名表現」になるのかなどを推定する。この推定は、規則に基づいた方法であってもよく、あるいは、機械学習を用いた方法であってもよい。前者の場合には、例えば、回答情報の解表現の種類を示す情報と、質問情報に含まれる語句を示す情報とを対応付けて有する情報を用いて、質問情報に、その語句を示す情報が含まれる場合に、その語句を示す情報に対応する解表現の回答情報になると判断してもよい。具体的には、解表現「人名」と、語句「誰」が対応付けられており、解表現「時間表現」と、語句「いつ」が対応付けられており、解表現「首都名」と、語句「首都はどこ」が対応付けられていてもよい。後者の場合には、例えば、あらかじめ用意された、質問情報と、その質問情報に対応する回答情報の解表現の種類を示す情報とを教師データとして機械学習を行い、その機械学習の結果を用いて、質問情報に対応する回答情報の解表現の種類を判断してもよい。機械学習の際には、教師データに含まれる質問情報に対して形態素解析を行い、その形態素解析で得られた形態素を素性として用いてもよい。なお、素性として用いるのは、形態素の表層（文字列そのもの）のみであってもよく、表層と品詞であってもよく、表層と品詞と活用形であってもよい。また、形態素を素性として用いるのではなく、質問情報の所定数のキャラクタ（文字）を素性として用いてもよい。この場合に、あらゆる文字列を素性として用いてもよく、所定数の文字列（例えば、３文字連続の文字列）を、１文字ずつずらしたものをすべて素性として用いてもよく、文頭から始まる文字列だけを素性として用いてもよく、文末で終わる文字列だけを素性として用いてもよい。例えば、質問情報から連続する２文字や３文字等を取得し、それらを素性として用いてもよい。機械学習としては、各種のアルゴリズムを用いることができる。このアルゴリズムの詳細については、［機械学習に関する説明］の欄で後述する。

【００９６】

次に、質問情報から用語を抽出する。この用語の抽出は、用語抽出部１３と同様の方法で行われうる。そして、文書取得手段２１と同様に、式（１）を用いて、コーパスから文書を抽出する。この文書の抽出により、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、質問情報が「日本の首都はどこですか」だとすると、例えば、「日本」「首都」が用語として抽出され、それらを含む文書群が取得されることになる。その後、前述のリランキングと同様の処理が行われる。なお、このリランキングの処理は、文書を絞り込むための処理であるので、行わなくてもよい。

【００９７】

次に、取得された文書、あるいは、リランキングで上位となった文書から、解を抽出する処理を行う。具体的には、それらの文書から、名詞、未知語連続を取り出して、それを解の候補とする。前述した非ファクトイド型の質問の異なり、ファクトイド型の質問の場合には、解が名詞、あるいは、名詞の連続となるため、このように名詞などを解の候補として抽出すればよいことになる。そして、その解の候補「ｃ」に対して、Ｓｃｏｒｅ（ｃ）を算出し、その値が大きいものを回答情報として選択して出力する。Ｓｃｏｒｅ（ｃ）は、次のようになる。

【００９８】

Ｓｃｏｒｅ（ｃ）＝$Score_{near}$（ｃ）＋$Score_{sem}$（ｃ）

$Score_{near}$（ｃ）は、解の候補とキーワードの近さに基づくスコアであり、$Score_{sem}$（ｃ）は、解表現の意味制約を満足しているかどうかに基づくスコアである。$Score_{near}$（ｃ）は、次式で与えられる。

【００９９】

【数４】

$$Score_{near}(c) = \sum_{t2 \in T3} w_{dr2}(t2) \log \frac{N}{2dist(c,t2) * df(t2)} \qquad (6)$$

$$T3 = \{t | t \in T, 2dist(c,t) \frac{df(t)}{N} \leq 1\} \qquad (7)$$

【０１００】

w_{dr2}（ｔ２）は、実験によって定められる関数であり、例えば、ｔ２が動詞であれば「０．５」となり、それ以外の品詞の用語であれば「１」となるものであってもよい。

【０１０１】

$Score_{sem}$（ｃ）は、解の候補の解表現の種類が、推定した解表現の種類と一致する場合に、正の値のスコアを与えて、そうでない場合に、スコアを与えない、あるいは、負の値のスコアを与えるという関数である。例えば、推定した解表現の種類（例えば、人名、地名等）と一致する解の候補にスコア（例えば、１０００）を与えてもよい。

【０１０２】

解の候補の解表現の種類が、推定した解表現の種類と一致するかどうかは、規則によって判断されてもよく、あるいは、機械学習によって判断されてもよい。前者の場合には、例えば、あらかじめ国名辞書、人名辞書、首都名辞書等を保持しておき、解の候補がいずれの辞書に記載されている単語であるのかを判断することによって、解の候補の解表現の種類を判断してもよい。また、後者の場合、すなわち、機械学習による場合には、あらかじめ用意された、解と、その解の解表現の種類を示す情報とを教師データとして機械学習を行い、その機械学習の結果を用いて、解の候補に対応する解表現の種類を判断してもよい。この機械学習による方法は、例えば、機械学習を用いた固有表現抽出技術などとして知られている方法である。

【０１０３】

このようにして、ファクトイド型の質問情報に対しても、回答情報を出力することができうる。なお、ファクトイド型の質問情報に対応する回答情報の取得・出力と、非ファクトイド型の質問情報に対応する回答情報の取得・出力において、類似の処理が行われることがある。例えば、質問情報からの用語の抽出や、コーパスからの文書の取得等である。したがって、ファクトイド型の質問情報に対応する回答情報の取得・出力と、非ファクトイド型の質問情報に対応する回答情報の取得・出力において、共通の構成要素（例えば、用語抽出部１３や、文書取得手段２１等）を用いて、処理を行うようにしてもよい。

【０１０４】

また、ファクトイド型の質問情報に対応する回答情報の抽出については、次の文献を参照されたい。

文献：村田真樹，内山将夫，白土保，井佐原均、「シリーズ型質問文に対して単純結合

法を利用した逓減的加点質問応答システム」、システム制御情報学会論文誌，Ｖｏｌ．２０，Ｎｏ．８，ｐ．１８－２６，２００７年

【０１０５】

なお、対応情報記憶部１４と、コーパス記憶部１５と、その他の情報が記憶される記録媒体とのうち、任意の２以上の記憶部や記録媒体は、同一の記録媒体によって実現されてもよく、あるいは、別々の記録媒体によって実現されてもよい。前者の場合には、例えば、対応情報を記憶している領域が対応情報記憶部１４となり、コーパスを記憶している領域がコーパス記憶部１５となる。

【０１０６】

また、コーパスや、用語、追加用語等は、厳密には、コーパスを示す情報や、用語を示す情報、追加用語を示す情報と記載すべきであるが、説明の便宜上、単にコーパスや、用語、追加用語等と呼ぶことにする。

【０１０７】

次に、本実施の形態による質問応答装置１の動作について、図３のフローチャートを用いて説明する。

（ステップＳ１０１）質問情報受付部１１は、質問情報を受け付けたかどうか判断する。そして、受け付けた場合には、ステップＳ１０２に進み、そうでない場合には、受け付けるまでステップＳ１０１の処理を繰り返す。

【０１０８】

（ステップＳ１０２）分類部１２は、質問情報受付部１１が受け付けた質問情報に対して、分類情報を付与する。

【０１０９】

（ステップＳ１０３）用語抽出部１３は、質問情報受付部１１が受け付けた質問情報から、用語を抽出する。この処理の詳細については、図４のフローチャートを用いて後述する。

【０１１０】

（ステップＳ１０４）回答情報取得部１６は、分類部１２によって付与された分類情報と、用語抽出部１３によって抽出された用語と、分類部１２によって付与された分類情報に、対応情報で対応付けられている追加用語と、コーパス記憶部１５で記憶されているコーパスとを用いて、質問情報受付部１１が受け付けた質問情報に対応する回答を示す回答情報を取得する。この処理の詳細については、図５のフローチャートを用いて後述する。

【０１１１】

（ステップＳ１０５）回答情報出力部１７は、回答情報取得部１６が取得した回答情報を出力する。そして、ステップＳ１０１に戻る。

なお、図３のフローチャートにおいて、電源オフや処理終了の割り込みにより処理は終了する。

【０１１２】

次に、図４は、図３のフローチャートにおける用語を抽出する処理（ステップＳ１０３の処理）の詳細を示すフローチャートである。なお、図４のフローチャートでは、用語抽出部１３が用語の抽出と共に、フォーカス表現の抽出も行う場合について説明する。また、図４のフローチャートでは、そのフォーカス表現の抽出を、手がかり句を用いて行う場合について説明する。

【０１１３】

（ステップＳ２０１）用語抽出部１３は、質問情報受付部１１が受け付けた質問情報を形態素解析する。

【０１１４】

（ステップＳ２０２）用語抽出部１３は、形態素解析の結果を用いて、あらかじめ決められている特定の品詞の用語を抽出する。

【０１１５】

（ステップＳ２０３）用語抽出部１３は、その抽出した特定の品詞の用語を、図示しな

い記録媒体において一時的に記憶する。

【０１１６】

（ステップＳ２０４）用語抽出部１３は、質問情報受付部１１が受け付けた質問情報に対して、分類部１２によって分類情報「定義質問」が付与されたかどうか判断する。そして、分類情報「定義質問」が付与された場合には、ステップＳ２０５に進み、そうでない場合には、図３のフローチャートに戻る。

【０１１７】

（ステップＳ２０５）用語抽出部１３は、あらかじめ図示しない記録媒体で保持されている、フォーカス表現を抽出するために用いられる手がかり句を読み出し、分類情報「定義質問」が付与された質問情報に、その手がかり句が含まれるかどうか判断する。そして、含まれる場合には、ステップＳ２０６に進み、そうでない場合には、図３のフローチャートに戻る。

【０１１８】

（ステップＳ２０６）用語抽出部１３は、質問情報に含まれる手がかり句と所定の関係にある単語を抽出する。この抽出した単語がフォーカス表現である。

【０１１９】

（ステップＳ２０７）用語抽出部１３は、抽出したフォーカス表現を図示しない記録媒体において一時的に記憶する。そして、図３のフローチャートに戻る。

【０１２０】

なお、図４のフローチャートにおいて、前述のように、フォーカス表現を機械学習やその他の方法を用いて抽出してもよいことは言うまでもない。また、後の回答情報の取得の処理において、フォーカス表現を用いない場合には、フォーカス表現の抽出の処理を行わなくてもよい。

【０１２１】

図５は、図３のフローチャートにおける回答情報の取得の処理（ステップＳ１０４の処理）の詳細を示すフローチャートである。なお、図５のフローチャートでは、回答情報取得部１６が、図２の構成である場合の処理について説明する。

【０１２２】

（ステップＳ３０１）文書取得手段２１は、前述の第２の式を用いて、用語抽出部１３が抽出した用語によって特徴付けられている程度の高い複数の文書を、コーパス記憶部１５で記憶されているコーパスから取得する。なお、この処理の詳細については、図６のフローチャートを用いて後述する。

【０１２３】

（ステップＳ３０２）追加用語取得手段２２は、分類部１２によって付与された分類情報に対応する１または２以上の追加用語を、対応情報記憶部１４で記憶されている対応情報から取得する。

【０１２４】

（ステップＳ３０３）算出手段２３は、用語抽出部１３が抽出した用語と、追加用語取得手段２２が取得した追加用語とを用いて、文書取得手段２１が取得した各文書に含まれる回答候補情報について、分類部１２によって付与された分類情報に応じた第１の式の値を算出する。この処理の詳細については、図７のフローチャートを用いて後述する。

【０１２５】

（ステップＳ３０４）回答情報選択手段２４は、算出手段２３が算出した値が他に比べて大きい値である複数の回答候補情報から、回答情報を選択する。そして、図３のフローチャートに戻る。なお、この処理の詳細については、図８のフローチャートを用いて後述する。

【０１２６】

図６は、図５のフローチャートにおける文書の取得の処理（ステップＳ３０１の処理）の詳細を示すフローチャートである。図６のフローチャートでは、文書取得手段２１が、前述の式（１）を第２の式として用いて、文書の取得をする場合について説明する。

【０１２７】
（ステップＳ４０１）文書取得手段２１は、カウンタｉを１に設定する。
（ステップＳ４０２）文書取得手段２１は、コーパス記憶部１５で記憶されているコーパスからｉ番目の文書を取得する。
【０１２８】
（ステップＳ４０３）文書取得手段２１は、用語抽出部１３が抽出した用語と、上記の式（１）と、コーパス記憶部１５で記憶されているコーパスとを用いて、第２の式の値、すなわち、式（１）の値を算出する。
【０１２９】
（ステップＳ４０４）文書取得手段２１は、算出した式（１）の値を、図示しない記録媒体に一時的に記憶する。なお、この記憶の際に、その式（１）の値に対応付けて、その値を算出した文書を識別する情報も記憶することが好適である。その文書を識別する情報は、例えば、カウンタの値であってもよく、コーパス記憶部１５で記憶されている文書の位置を示すポインタであってもよく、あるいは、その他の文書の識別情報であってもよい。
（ステップＳ４０５）文書取得手段２１は、カウンタｉを１だけインクリメントする。
【０１３０】
（ステップＳ４０６）文書取得手段２１は、コーパス記憶部１５で記憶されているコーパスに、ｉ番目の文書が存在するかどうか判断する。そして、存在する場合には、ステップＳ４０２に戻り、そうでない場合には、ステップＳ４０７に進む。
【０１３１】
（ステップＳ４０７）文書取得手段２１は、ステップＳ４０４で一時的に記憶した式（１）の値をソートする。
【０１３２】
（ステップＳ４０８）文書取得手段２１は、ソート結果において、式（１）の値が他に比べて大きい複数の文書を選択する。そして、図５のフローチャートに戻る。なお、文書取得手段２１は、前述のように、式（１）の値がしきい値以上の文書を選択してもよく、式（１）の値が大きい方から所定数、あるいは所定割合の文書を選択してもよい。また、この文書の選択は、前述のように、文書を識別する情報の取得であってもよく、コーパス記憶部１５からの文書の情報そのものの取得であってもよい。
【０１３３】
図７は、図５のフローチャートにおける第１の式の算出の処理（ステップＳ３０３の処理）の詳細を示すフローチャートである。図７のフローチャートでは、前述の式（２）を用いて第１の式の値を算出する場合について説明する。また、質問情報に付与された分類情報が「程度質問」であり、かつ、回答候補情報に数表現がある場合には、式（２）の値が１．１倍されるものとする。また、質問情報に付与された分類情報が「定義質問」であり、かつ、回答候補情報にフォーカス表現が含まれる場合には、式（２）の値が１．１倍されるものとする。また、また、質問情報に付与された分類情報が「定義質問」であり、かつ、回答候補情報にフォーカス表現が含まれ、かつ、回答候補情報においてフォーカス表現が連体修飾節で修飾されている場合には、式（２）の値が１．１倍されるものとする。
【０１３４】
（ステップＳ５０１）算出手段２３は、カウンタｉを１に設定する。
（ステップＳ５０２）算出手段２３は、カウンタｊを１に設定する。
【０１３５】
（ステップＳ５０３）算出手段２３は、文書取得手段２１が取得したｉ番目の文書において、ｊ番目の回答候補情報を特定する。算出手段２３は、例えば、ｊ番目の回答候補情報を取得することによって、その特定を行ってもよく、ｊ番目の回答候補情報が記憶されている位置を示すポインタ等を取得することによって、その特定を行ってもよく、結果として、後の処理で特定された回答候補情報を用いることができるのであれば、その特定の

方法を問わない。

【０１３６】

（ステップＳ５０４）算出手段２３は、用語抽出部１３が抽出した用語と、追加用語取得手段２２が取得した追加用語と、ステップＳ５０３で特定した回答候補情報と、上記の式（２）、（３）と、コーパス記憶部１５で記憶されているコーパスとを用いて、式（２）の値を算出する。なお、質問情報に付与された分類情報が「程度質問」「定義質問」でない場合には、この式（２）の値が、第１の式の値となる。

【０１３７】

（ステップＳ５０５）算出手段２３は、算出した式（２）の値を、図示しない記録媒体に一時的に記憶する。なお、この記憶の際に、その式（２）の値に対応付けて、その値を算出した回答候補情報を識別する情報も記憶することが好適である。その回答候補情報を識別する情報は、例えば、カウンタｉ，ｊの値であってもよく、その回答候補情報の含まれる文書を識別する情報と、その文書における回答候補情報の位置を示すポインタであってもよく、回答候補情報そのものであってもよく、あるいは、その他の識別情報であってもよい。

【０１３８】

（ステップＳ５０６）算出手段２３は、質問情報受付部１１が受け付けた質問情報に付与された分類情報が程度質問であるかどうか判断する。そして、程度質問である場合には、ステップＳ５０７に進み、そうでない場合には、ステップＳ５１３に進む。

【０１３９】

（ステップＳ５０７）算出手段２３は、ｉ番目の文書におけるｊ番目の回答候補情報に、数表現が含まれるかどうか判断する。そして、数表現が含まれる場合には、ステップＳ５０８に進み、そうでない場合には、ステップＳ５０９に進む。

【０１４０】

（ステップＳ５０８）算出手段２３は、ステップＳ５０５で一時的に記憶した式（２）の値を１．１倍して、上書きで蓄積する。その上書き後の値が、第１の式の値となる。

（ステップＳ５０９）算出手段２３は、カウンタｊを１だけインクリメントする。

【０１４１】

（ステップＳ５１０）算出手段２３は、ｉ番目の文書にｊ番目の回答候補情報が存在するかどうか判断する。そして、存在する場合には、ステップＳ５０３に戻り、存在しない場合には、ステップＳ５１１に進む。

（ステップＳ５１１）算出手段２３は、カウンタｉを１だけインクリメントする。

【０１４２】

（ステップＳ５１２）算出手段２３は、文書取得手段２１によって取得されたｉ番目の文書が存在するかどうか判断する。そして、存在する場合には、ステップＳ５０２に戻り、存在しない場合には、図５のフローチャートに戻る。

【０１４３】

（ステップＳ５１３）算出手段２３は、質問情報受付部１１が受け付けた質問情報に付与された分類情報が定義質問であるかどうか判断する。そして、定義質問である場合には、ステップＳ５１４に進み、そうでない場合には、ステップＳ５０９に進む。

【０１４４】

（ステップＳ５１４）算出手段２３は、ｉ番目の文書におけるｊ番目の回答候補情報に、用語抽出部１３が抽出したフォーカス表現が存在するかどうか判断する。そして、存在する場合には、ステップＳ５１５に進み、存在しない場合（フォーカス表現の抽出が行われていない場合を含む）には、ステップＳ５０９に進む。

【０１４５】

（ステップＳ５１５）算出手段２３は、ステップＳ５０５で一時的に記憶した式（２）の値を１．１倍して、上書きで蓄積する。これより後にその値の上書きが行われない場合には、その上書き後の値が、第１の式の値となる。

【０１４６】

（ステップＳ５１６）算出手段２３は、ｉ番目の文書におけるｊ番目の回答候補情報に含まれるフォーカス表現が、連体修飾節によって修飾されているかどうか判断する。そして、連体修飾節によって修飾されている場合には、ステップＳ５１７に進み、そうでない場合には、ステップＳ５０９に進む。

【０１４７】

（ステップＳ５１７）算出手段２３は、算出手段２３は、ステップＳ５１６において上書きで蓄積した値をさらに１．１倍して、上書きで蓄積する。その上書き後の値が、第１の式の値となる。

【０１４８】

（ステップＳ５１８）算出手段２３は、ｉ番目の文書におけるｊ番目の回答候補情報を、フォーカス表現を修飾する連体修飾節に置き換える。そして、ステップＳ５０９に進む。

【０１４９】

なお、図７のフローチャートにおけるステップＳ５１８の処理は、算出手段２３によって行われてもよく、あるいは、回答情報選択手段２４によって行われてもよい。また、図７のフローチャートでは、質問情報が「程度質問」「定義質問」に分類された場合についてのみ、第１の式の値を、式（２）から変更する場合について説明したが、質問情報がその他の分類に分類された場合についても、第１の式の値を、式（２）から変更するようにしてもよい。また、その変更の程度が、「１．１倍」である場合について説明したが、そうでなくてもよい。

【０１５０】

図８は、図５のフローチャートにおける回答情報の選択の処理（ステップＳ３０４の処理）の詳細を示すフローチャートである。図８のフローチャートにおいて、回答情報選択手段２４は、回答候補情報の選択と共に、選択された回答候補情報に含まれる回答情報の特定の処理を行うものとする。その回答情報の特定は、機械学習を用いて行われるものとする。なお、ステップＳ６０１からの処理が実行される前に、あらかじめ機械学習の処理が行われているものとする。

【０１５１】

（ステップＳ６０１）回答情報選択手段２４は、算出手段２３によって算出された値を用いて、回答候補情報をソートする。例えば、回答情報選択手段２４は、回答候補情報が、その回答候補情報に対応する値（算出手段２３によって算出された値）の降順となるようにソートする。

【０１５２】

（ステップＳ６０２）回答情報選択手段２４は、ソート結果において、算出手段２３によって算出された値が他に比べて大きい１以上の回答候補情報を選択する。なお、回答情報選択手段２４は、前述のように、算出手段２３の算出した値がしきい値以上の回答候補情報を選択してもよく、算出手段２３の算出した値が大きい方から所定数、あるいは所定割合の回答候補情報を選択してもよい。

【０１５３】

（ステップＳ６０３）回答情報選択手段２４は、カウンタｉを１に設定する。

（ステップＳ６０４）回答情報選択手段２４は、カウンタｊを１に設定する。

【０１５４】

（ステップＳ６０５）回答情報選択手段２４は、ｉ番目の回答候補情報において、ｊ番目の部分を特定する。特定される部分の単位は、あらかじめ決まっていてもよい。例えば、特定される部分が文単位である場合には、回答情報選択手段２４は、ｉ番目の回答候補情報において、ｊ番目の文を特定する。また、例えば、特定される部分がパラグラフ単位である場合には、回答情報選択手段２４は、ｉ番目の回答候補情報において、ｊ番目のパラグラフを特定する。

【０１５５】

（ステップＳ６０６）回答情報選択手段２４は、ステップＳ６０５で特定した部分が回

答情報であるかどうかを、機械学習の結果を用いて判断する。そして、回答情報であると判断した場合には、ステップＳ６０７に進み、そうでない場合には、ステップＳ６０８に進む。

【０１５６】

（ステップＳ６０７）回答情報選択手段２４は、回答情報であると判断した部分を、回答情報として、図示しない記録媒体において一時的に記憶する。

（ステップＳ６０８）回答情報選択手段２４は、カウンタｊを１だけインクリメントする。

【０１５７】

（ステップＳ６０９）回答情報選択手段２４は、ｉ番目の回答候補情報に、ｊ番目の部分が存在するかどうか判断する。そして、存在する場合には、ステップＳ６０５に戻り、存在しない場合には、ステップＳ６１０に進む。

（ステップＳ６１０）回答情報選択手段２４は、カウンタｉを１だけインクリメントする。

【０１５８】

（ステップＳ６１１）回答情報選択手段２４は、算出手段２３が値を算出した回答候補情報に、ｉ番目の回答候補情報が存在するかどうか判断する。そして、存在する場合には、ステップＳ６０４に戻り、そうでない場合には、図５のフローチャートに戻る。

【０１５９】

なお、図８のフローチャートにおいて、選択された回答候補情報の部分を回答情報とする場合について説明したが、その処理を行わなくてもよい。例えば、ステップＳ６０２で選択された回答候補情報を、回答情報としてもよい。また、ステップＳ６０２で選択された回答候補情報、あるいは、ステップＳ６０２の処理を行わない回答候補情報から、機械学習の結果を用いて回答候補情報を選択して、その選択した回答候補情報を回答情報としてもよい。その場合には、カウンタｊを用いずに（すなわち、ｉ番目の回答候補情報におけるｊ番目の部分の特定を行わずに）、機械学習の結果を用いて、ｉ番目の回答候補情報が回答情報であるかどうかの判断を行ってもよい。また、図８のフローチャートにおいて、ステップＳ６０１、Ｓ６０２の処理を行わずに、機械学習の結果を用いた回答情報の選択のみの処理（ステップＳ６０３～Ｓ６１１）を行うようにしてもよい。なお、図８のフローチャートでステップＳ６０２の処理を行わない場合であっても、機械学習の際に、第１の式の値を教師データとして用いることによって、回答情報取得部１６は、間接的に、追加用語、用語抽出部１３が抽出した用語、分類情報に応じた式（第１の式）とを用いて回答情報を取得することになる。また、図７のフローチャートにおいて、回答候補情報が、フォーカス表現を修飾する連体修飾節に置き換えられた場合（ステップＳ５１８の処理が実行された場合）には、ステップＳ６０５において、回答候補情報そのものを特定するようにしてもよい。

【０１６０】

また、図８のフローチャートでは、１以上の回答情報を選択する場合について説明したが、１個の回答情報を選択するようにしてもよい。例えば、ステップＳ６０２において、１個の回答候補緒情報のみを選択するようにしてもよく、あるいは、ステップＳ６０６における機械学習の結果を用いた判断において、確からしさ（確信度）の最も高い回答情報を選択するようにしてもよい。

【０１６１】

次に、本実施の形態による質問応答装置１の動作について、具体例を用いて説明する。

この具体例において、本実施の形態による質問応答装置１は、スタンドアロンのＰＣ（Ｐｅｒｓｏｎａｌ　Ｃｏｍｐｕｔｅｒ）であるとする。

【０１６２】

また、この具体例において、対応情報記憶部１４では、図９で示される対応情報が記憶されているものとする。図９の対応情報において、分類情報と、追加用語とが対応付けられている。なお、この具体例において、分類部１２は、質問情報を、定義質問、理由質問

、方法質問、程度質問、変化質問、経緯質問の６種類に分類するものとするが、図９で示されるように、追加用語と対応していない分類情報が存在してもよい。また、この具体例において、質問情報の分類は、機械学習を用いて行うものとする。

【０１６３】

まず、ユーザが、質問応答装置１のマウスやキーボード等を操作することにより、図１０で示されるように、質問入力画面を表示させたとする。そして、ユーザが、キーボード等を操作することによって、質問情報「世界遺産は、どのようにして決まるのですか。」を入力し、「ＯＫ」ボタンをクリックしたとする。すると、質問情報受付部１１は、質問情報「世界遺産は、どのようにして決まるのですか。」を受け付け（ステップＳ１０１）、その質問情報を分類部１２と、用語抽出部１３とに渡す。

【０１６４】

分類部１２では、あらかじめ、分類に関する機械学習を行っているものとする。そして、質問情報「世界遺産は、どのようにして決まるのですか。」を質問情報受付部１１から受け付けると、その学習結果を用いて、質問情報を分類する。この場合には、分類部１２は、その質問情報に分類情報「方法質問」を付与したとする（ステップＳ１０２）。分類部１２は、その分類結果（付与された分類情報）を用語抽出部１３と、回答情報取得部１６に渡す。

【０１６５】

用語抽出部１３は、質問情報を質問情報受付部１１から受け取り、分類情報を分類部１２から受け取ると、用語を抽出する処理を実行する（ステップＳ１０３）。具体的には、用語抽出部１３は、質問情報「世界遺産は、どのようにして決まるのですか。」を形態素解析し（ステップＳ２０１）、抽出する対象にあらかじめ決められている品詞を抽出する（ステップＳ２０２）。ここでは、名詞と動詞を抽出するように決められているものとする。すると、用語抽出部１３は、形態素解析された結果から、名詞「世界遺産」と、動詞「決まる」とを抽出し（ステップＳ２０２）、図示しない記録媒体において一時的に記憶する（ステップＳ２０３）。なお、この具体例では、複合名詞「世界遺産」を抽出する場合について説明するが、名詞「世界」と、名詞「遺産」とを抽出してもよく、また、複合名詞「世界遺産」と共に、名詞「世界」と、名詞「遺産」を抽出してもよく、用語抽出部１３が用語を抽出する方法としては、各種の方法が存在することになる。また、複合名詞を抽出する方法として、例えば、連続した名詞を複合名詞として抽出する方法を用いてもよく、前述した固有表現抽出と同様に、人手のルールや機械学習を用いて抽出する方法を用いてもよく、前述したＴｅｒｍＥｘｔｒａｃｔを用いて抽出する方法を用いてもよく、複合名詞を抽出することができるのであれば、その方法は限定されない。

【０１６６】

また、用語抽出部１３は、分類部１２から受け取った分類情報が「定義質問」であるかどうか判断する。この場合には、分類情報は「方法質問」であるため、用語抽出部１３は、分類情報が「定義質問」ではないと判断し（ステップＳ２０４）、その後のフォーカス表現を抽出する処理は行わない。

【０１６７】

その後、回答情報取得部１６は、回答情報を取得する処理を実行する（ステップＳ１０４）。その回答情報を取得する処理において、まず、文書取得手段２１は、コーパス記憶部１５から文書を取得する処理を実行する（ステップＳ３０１）。具体的には、文書取得手段２１は、コーパス記憶部１５で記憶されている１番目の文書を読み出す（ステップＳ４０１，Ｓ４０２）。ここで、その１番目の文書を識別する文書ＩＤは、「Ｄ００１」であったとする。次に、文書取得手段２１は、前述の式（１）が記憶されている図示しない記録媒体から式（１）を読み出し、また、用語抽出部１３が一時的に記憶した用語「世界遺産」「決まる」を読み出し、それらを用いて、式（１）の値を算出する（ステップＳ４０３）。ここでは、式（１）の値が「２．３」であったとする。そして、文書取得手段２１は、文書ＩＤ「Ｄ００１」に対応付けて、その式（１）の値「２．３」を図示しない記録媒体で一時的に記憶する（ステップＳ４０４）。図１１の１番目のレコードは、そのよ

うにして記憶されたものである。同様の処理を、文書取得手段２１は、２番目の文書、３番目の文書、…について、順次、実行する（ステップＳ４０５，Ｓ４０６，Ｓ４０２～Ｓ４０４）。

【０１６８】

すべての文書について、式（１）の値を算出した後に、文書取得手段２１は、算出した値を降順となるようにソートする（ステップＳ４０７）。そして、そのソートした結果において、値の大きいものから３００個の文書ＩＤを取得して図示しない記録媒体に蓄積する（ステップＳ４０８）。このようにして、文書取得手段２１による文書の取得が行われる。ここでは、文書ＩＤ「Ｄ００２」「Ｄ００３」…が選択されたものとする。

【０１６９】

次に、追加用語取得手段２２は、分類部１２から受け取った分類情報「方法質問」をキーとして、図９で示される対応情報を検索し、検索された「方法質問」に対応付けられている追加用語「方法」「手順」「ことにより」を読み出して、図示しない記録媒体に蓄積する（ステップＳ３０２）。

【０１７０】

文書取得手段２１による文書の取得と、追加用語取得手段２２による追加用語の取得との後に、算出手段２３は、第１の式の値を算出する処理を実行する（ステップＳ３０３）。ここで、算出手段２３は、文書取得手段２１が取得した文書において、１パラグラフを回答候補情報として、第１の式の値を算出するものとする。また、文書取得手段２１が取得した、文書ＩＤ「Ｄ００２」で識別される文書は、図１２で示されるものであったとする。

【０１７１】

すると、算出手段２３は、まず、１番目の文書である、図１２で示される文書の１番目のパラグラフ、すなわち、「世界遺産とは、……、もつものである。」を回答候補情報として特定する（ステップＳ５０１～Ｓ５０３）。そして、算出手段２３は、用語抽出部１３が蓄積した用語「世界遺産」「決まる」と、追加用語取得手段２２が蓄積した追加用語「方法」「手順」「ことにより」とを読み出し、図示しない記録媒体で記憶されている式（３）も読み出し、それらとコーパスとを用いて、Ｔ３のセットを算出する。その後、算出したＴ３のセットと、用語「世界遺産」「決まる」や追加用語「方法」「手順」「ことにより」、コーパス、図示しない記録媒体から読み出した式（２）などを用いて、式（２）の値を算出する（ステップＳ５０４）。ここでは、２．６となったものとする。すると、算出手段２３は、その値を、回答候補情報を識別する情報に対応付けて図示しない記録媒体に蓄積する（ステップＳ５０５）。図１３の１番目のレコードは、そのようにして蓄積された第１の式のスコアと、回答候補情報の識別情報とを対応付けて有する情報である。回答候補情報の識別情報としては、文書ＩＤと、文書におけるパラグラフ番号とが用いられている。パラグラフ番号は、文書の先頭から数えたパラグラフの数を示す値である。

【０１７２】

この場合には、分類情報は「方法質問」であって、「程度質問」や「定義質問」ではないため、回答候補情報の特定と、式（２）の値の算出と、その蓄積とが順次、行われることになる（ステップＳ５０９，Ｓ５１０，Ｓ５０３～Ｓ５０５）。また、１番目の文書について終了すれば、順次、２番目、３番目、…の文書についても、同様に、その文書に含まれるパラグラフごとの式（２）の値が算出され、蓄積されていく（ステップＳ５０２～Ｓ５０５，Ｓ５０９～Ｓ５１２）。

【０１７３】

第１の式の値を算出する一連の処理が終了すると、回答情報選択手段２４は、回答情報を選択する処理を実行する（ステップＳ３０４）。具体的には、回答情報選択手段２４は、図１３で示される算出結果を図示しない記録媒体から読み出し、第１の式のスコアの降順となるように各レコードをソートする（ステップＳ６０１）。そして、第１の式のスコアの最大値（ここでは、９．７であったとする）に０．９を掛けた値（＝８．７３）以上の第１の式のスコアを有する回答候補情報を選択する（ステップＳ６０２）。ここで、選

択された回答候補情報は、文書ＩＤ「Ｄ００２」と、パラグラフ番号「３」で識別される回答候補情報のみであったとする。また、回答情報選択手段２４は、回答候補情報に含まれる文ごとに、その文が回答情報として、適切であるかどうかを、機械学習によって判断するものとする。その判断のための機械学習は、あらかじめ、行われているものとする。

【０１７４】

すると、回答情報選択手段２４は、選択された１番目の回答候補情報である、図１２で示される段落「登録の手順としては、……」から、１番目の文を特定する（ステップＳ６０３～Ｓ６０５）。そして、回答情報選択手段２４は、機械学習の結果を用いて、その特定した文「登録の手順としては、……専門家団体が評価する。」が回答情報であるかどうか判断する。ここでは、その文が回答情報であると判断されたとする（ステップＳ６０６）。すると、回答情報選択手段２４は、その文を回答情報として、図示しない記録媒体に蓄積し（ステップＳ６０７）、次の２番目の文「そして、その評価結果に基づいて、……決定する。」を特定し、その文が回答情報であるかどうか判断する（ステップＳ６０８，Ｓ６０９，Ｓ６０５，Ｓ６０６）。この場合には、回答情報ではないと判断されたとする。そして、回答情報を選択する一連の処理は、終了となる。

【０１７５】

その後、回答情報出力部１７は、回答情報選択手段２４が蓄積した回答情報を図示しない記録媒体から読み出し、その回答情報を質問応答装置１のディスプレイに表示する（ステップＳ１０５）。図１４は、そのようにして表示された回答情報を示す図である。質問を入力したユーザは、この表示を見ることによって、質問への回答を知ることができうる。

【０１７６】

なお、回答情報を質問応答装置１のディスプレイに表示する際に、その回答情報が含まれる文書全体をディスプレイに表示し、その上で、回答情報と、回答情報以外の文書の箇所とを区別可能なように表示してもよい。区別可能なように表示するとは、例えば、回答情報と、回答情報以外の文書の箇所との表示の色を変更することであってもよく、表示のフォントの大きさやフォントの種類を変更することであってもよく、回答情報の箇所にのみ下線を付与することであってもよく、回答情報の箇所のみ、枠囲みや網掛け等を行うことであってもよく、回答情報以外の文書の箇所を、回答情報よりも薄く表示することであってもよい。

また、この具体例において用いた数値等は、説明のために設定した値であって、実際のデータを解析することによって算出したものではない。

【０１７７】

また、この具体例においては、機械学習によって分類を行う場合について説明したが、前述のように、分類対応情報を用いて分類を行ってもよいことは言うまでもない。その場合に用いる分類情報は、例えば、図１５で示されるものであってもよい。

【０１７８】

［実験例］

次に、実験例について説明する。この実験において、１００個の非ファクトイド型の質問情報を用いた。その質問情報は、ＱＡＣ－４の主催者によって生成されたものであり、ターゲット文書を用いないで生成された自然な質問である。また、ＱＡＣ－４の主催者は、各質問情報に対して、４個以下の回答情報を評価した。結果は、図１６に示すとおりである。方法１は、回答候補情報として、１パラグラフを用いた方法である。方法２は、回答候補情報として、１パラグラフと、連続した２パラグラフと、連続した３パラグラフとを用いた方法である。Ａ，Ｂ，Ｃ，Ｄは、評価基準である。Ａは、回答情報が、質問情報に対して主催者が用意した回答と同内容を記述している場合である。追加的な内容を含んでいたとしても、内容が変わらないものについては、Ａと評価した。Ｂは、回答情報が、質問情報に対して主催者が用意した回答と類似する内容を含むが、全体として異なる内容も含んでいる場合である。Ｃは、回答情報が、質問情報に対して主催者が用意した回答と同じ内容の一部を含む場合である。Ｄは、回答情報が、質問情報に対して主催者が用意し

た回答と同じ内容を含まない場合である。図１６のテーブルにおいて、Ａ，Ｂ，Ｃ，Ｄの値は、回答情報がＡ，Ｂ，Ｃ，Ｄに属する質問情報の数である。「正解」は、回答情報がＡ，Ｂ，Ｃのいずれかに属する場合の質問情報の数である。この正解の評価基準は、ＮＴＣＩＲ－６　ＱＡＣ－４でも公式に用いられたものである。

【０１７９】

この図１６で示される結果から、次のようなことが分かる。

方法１の方が方法２よりも、Ａの評価において、より高いスコアを得ている。このことから、方法１が方法２よりも、完全な正解をより正確に抽出していることが分かる。

【０１８０】

「正解」の評価において、方法２の正解率は０．７７であり、方法２の方が方法１よりも、より高いスコアを得ている。このことから、方法２の方が方法１よりも、部分的な正解をより多く抽出している傾向が分かる。したがって、完全な正解を抽出したい場合には、方法１を用いればよく、より多くの正解（部分的な正解を含む）を抽出したい場合には、方法２を用いればよい。

【０１８１】

なお、ＮＴＣＩＲ－６　ＱＡＣ－４に参加した全８チーム中、本実施の形態による質問応答装置１を用いた方法２の正解率「０．７７」は、もっとも高い値であった。このことから、本実施の形態による質問応答装置１が、他の参加チームの質問応答装置と比較して、最も効果的に回答情報の取得を行うことができたことが分かる。

【０１８２】

ここで、本実施の形態による質問応答装置１によって実際に取得した回答情報の例について簡単に説明する。この例では、方法１を用いて、Ａと評価された例を示す。

【０１８３】

質問情報：受精卵診断は、どういう場合に行われるのか？

回答情報：主に遺伝病の子供が生まれるのを防ぐ

ＱＡＣ－４主催者が用意した回答例１：主に遺伝病

ＱＡＣ－４主催者が用意した回答例２：主に遺伝病の子供が生まれる可能性が高い場合

【０１８４】

なお、この実験例の結果を得る際に用いた質問応答装置１では、式（１）～（３）を用いた回答情報の取得を行った。そして、式（１）を用いて３００の文書を取得し、その文書に含まれる回答候補情報に対して算出した式（２）の値の最大値に０．９を掛けた値以上を有する回答候補情報を、回答情報として取得して出力した。したがって、図８のフローチャートのステップＳ６０４～Ｓ６０９で説明した、機械学習を用いて回答候補情報から回答情報を取得する処理は行っていない。

【０１８５】

以上のように、本実施の形態による質問応答装置１によれば、非ファクトイド型の質問情報に対応する回答情報を適切に抽出して出力することができる。また、実験例の結果で示されるように、本実施の形態による質問応答装置１の手法が、最も高い正解率の得られる手法であり、本実施の形態による質問応答装置１が、他の従来の質問応答装置に比べて高性能であることが分かる。

【０１８６】

なお、本実施の形態において、質問情報、回答情報、コーパスの言語は問わない。例えば、日本語、英語、フランス語、ロシア語、中国語、スペイン語等であってもよい。また、質問情報と、コーパスとの言語が異なっていてもよい。例えば、質問情報が日本語で、コーパスが英語である場合などである。その場合には、例えば、質問情報をコーパスの言語に翻訳した上で、処理を行ってもよく、コーパスを質問情報の言語に翻訳した上で処理を行ってもよい。一般に、前者の方が、翻訳量が少なくなると考えられ、好適である。その翻訳は、例えば、既存の機械翻訳の手法を用いてもよい。機械翻訳としては、例えば、統計的な機械翻訳等を用いることができうる。

【０１８７】

［機械学習に関する説明］
ここで、機械学習について説明する。機械学習の手法は、問題－解の組のセットを多く用意し、そのセットを用いて学習を行なうことによって、どういう問題のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときも解を推測できるようにする方法である。例えば、次の文献を参照されたい。
【０１８８】
文献：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年（ｈｔｔｐ：／／ｗｗｗ２．ｎｉｃｔ．ｇｏ．ｊｐ／ｊｔ／ａ１３２／ｍｅｍｂｅｒｓ／ｍｕｒａｔａ／ｐｓ／ｒｋ１－ｓｉｒｙｏｕ．ｐｄｆ）
【０１８９】
文献：村田真樹，馬青，内元清貴，井佐原均、「サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳」，電子情報通信学会言語理解とコミュニケーション研究会　ＮＬＣ２０００－７８，２００１年
【０１９０】
文献：村田真樹，内山将夫，内元清貴，馬青，井佐原均、「ＮＳＥＶＡＬ２Ｊ辞書タスクでのＣＲＬの取り組み」、電子情報通信学会言語理解とコミュニケーション研究会　ＮＬＣ２００１－４０，２００１年
【０１９１】
機械学習アルゴリズムを動作させるために、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「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。
【０１９２】
すなわち、機械学習の手法は、素性の集合－解の組のセットを多く用意し、そのセットを用いて学習を行なうことによって、どういう素性の集合のときにどういう解になるかを学習し、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て、その素性に対応する解を推測する方法である。なお、ここで、「解」とは、例えば、前述の回答情報であるかどうかや、分類情報などである。
【０１９３】
機械学習の手法として、例えば、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ことができる。なお、以下の説明では、文書を分類する場合（問題－解のセットが、文－分類である場合）の機械学習について主に説明するが、それ以外の機械学習についても、同様に適用可能であることは言うまでもない。
【０１９４】
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解（分類）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。
【０１９５】
シンプルベイズ法は、ベイズの定理にもとづいて各解（分類）の確率を推定し、その確率値が最も大きい解を、求める解とする方法である。
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、次式で与えられる。
【０１９６】

【数５】

$$p(a|b) = \frac{p(a)}{p(b)}\, p(b|a)$$
$$\cong \frac{\tilde{p}(a)}{p(b)} \prod_i \tilde{p}(f_i|a)$$

【０１９７】

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、上記の２行目の式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、スムージングを行う。ここでは、次式を用いてスムージングを行ったものを用いる。

【０１９８】

【数６】

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

【０１９９】

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。

なお、スムージングは、上記式を用いた方法に限られるものではなく、その他の方法を用いてもよいことは言うまでもない。

【０２００】

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。

【０２０１】

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、次式によって与えられる。

【数７】

$$p(a|b) = p(a|f_{\max})$$

【０２０２】

ただし、ｆ$_{max}$は、次式によって与えられる。

【数８】

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

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。

【０２０３】

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、以下の所定の条件式を満足しながらエントロピーを意味する式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。

【０２０４】
所定の条件式は、次式で与えられる。
【数９】

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

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

【０２０５】
また、エントロピーを意味する式は、次式で与えられる。
【数１０】

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

【０２０６】
ただし、Ａ、Ｂは分類と文脈の集合を意味する。また、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。
【０２０７】
上記の条件式は、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、以下の文献を参照されたい。
【０２０８】
文献：Ｅｒｉｃ Ｓｖｅｎ Ｒｉｓｔａｄ，「Ｍａｘｉｍｕｍ Ｅｎｔｒｏｐｙ Ｍｏｄｅｌｉｎｇ ｆｏｒ Ｎａｔｕｒａｌ Ｌａｎｇｕａｇｅ」，（ＡＣＬ／ＥＡＣＬ Ｔｕｔｏｒｉａｌ Ｐｒｏｇｒａｍ，Ｍａｄｒｉｄ，１９９７年
【０２０９】
文献：Ｅｒｉｃ Ｓｖｅｎ Ｒｉｓｔａｄ，「Ｍａｘｉｍｕｍ Ｅｎｔｒｏｐｙ Ｍｏｄｅｌｉｎｇ Ｔｏｏｌｋｉｔ，Ｒｅｌｅａｓｅ１．６ｂｅｔａ」，（ｈｔｔｐ：／／ｗｗｗ．ｍｎｅｍｏｎｉｃ．ｃｏｍ／ｓｏｆｔｗａｒｅ／ｍｅｍｔ），１９９８年
【０２１０】
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。
【０２１１】
図１７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１７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１７（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１７（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。
【０２１２】
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１７（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。
【０２１３】
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型にする拡張（カーネル関数の導入）がなされたものが用いられる。

【０２１４】
この拡張された方法は、以下の識別関数（ｆ（ｘ））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。
【数１１】

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\right) \qquad \text{(M1)}$$

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

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

【０２１５】
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（Ｍ３）と式（Ｍ４）の制約のもと、式（Ｍ２）を最大にする場合のものである。
【０２１６】
【数１２】

$$L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \qquad \text{(M2)}$$

$$0\le\alpha_i\le C \qquad (i=1,\ldots,l) \qquad \text{(M3)}$$

$$\sum_{i=1}^{l}\alpha_i y_i=0 \qquad \text{(M4)}$$

【０２１７】
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、例えば、以下の多項式のものを用いる。
Ｋ（ｘ，ｙ）＝（ｘ・ｙ＋１）ｄ
【０２１８】
ここで、Ｃ、ｄは実験的に設定される定数である。例えば、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（Ｍ１）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。
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、次の文献を参照されたい。
【０２１９】
文献：Ｎｅｌｌｏ Ｃｒｉｓｔｉａｎｉｎｉ，Ｊｏｈｎ Ｓｈａｗｅ－Ｔａｙｌｏｒ，「Ａｎ Ｉｎｔｒｏｄｕｃｔｉｏｎ ｔｏ Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅｓ ａｎｄ ｏｔｈｅｒ ｋｅｒｎｅｌ－ｂａｓｅｄ ｌｅａｒｎｉｎｇ ｍｅｔｈｏｄｓ」，Ｃａｍｂｒｉｄｇｅ Ｕｎｉｖｅｒｓｉｔｙ Ｐｒｅｓｓ，２０００年
【０２２０】
文献：Ｔａｋｕ Ｋｕｄｏｈ，「Ｔｉｎｙｓｖｍ：Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ ｍａｃｈｉｎｅｓ」，（ｈｔｔｐ：／／ｃｌ．ａｉｓｔｎａｒａ．ａｃ．ｊｐ／ｔａｋｕ－

ｋｕ／／ｓｏｆｔｗａｒｅ／Ｔｉｎｙ　ＳＶＭ／ｉｎｄｅｘ．ｈｔｍｌ），２０００年
【０２２１】
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。
【０２２２】
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、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。
【０２２３】
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。推定するべき問題が、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、ある解くべき問題が、「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、その解くべき問題の分類先は、ａと推定する。
【０２２４】
図示しない解推定手段が推定する、解くべき問題についての、どのような解（分類先）になりやすいかの度合いの求め方は、図示しない機械学習手段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。
【０２２５】
例えば、機械学習手段が、機械学習の手法としてｋ近傍法を用いる場合、機械学習手段は、教師データの事例同士で、その事例から抽出された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）にもとづく事例同士の類似度を定義して、前記定義した類似度と事例とを学習結果情報として学習結果記憶手段に記憶しておく。
【０２２６】
そして、解推定手段は、解くべき問題の素性（文章群属性情報）が抽出されたときに、学習結果記憶手段において定義された類似度と事例を参照して、抽出された解くべき問題の素性について、その解くべき問題の素性の類似度が高い順にｋ個の事例を学習結果記憶手段の事例から選択し、選択したｋ個の事例での多数決によって決まった分類先を、解くべき問題の分類先（解）として推定する。すなわち、解推定手段では、解くべき問題についての、どのような解（分類先）になりやすいかの度合いを、選択したｋ個の事例での多数決の票数とする。
【０２２７】
また、機械学習手法として、シンプルベイズ法を用いる場合には、図示しない機械学習手段は、教師データの事例について、前記事例の解と素性の集合との組を学習結果情報として学習結果記憶手段に記憶する。そして、解推定手段は、解くべき問題の素性が抽出されたときに、学習結果記憶手段の学習結果情報の解と素性の集合との組をもとに、ベイズの定理にもとづいて、解くべき問題の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、その解くべき問題の素性の分類（解）と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。
【０２２８】
また、機械学習手法として決定リスト法を用いる場合には、図示しない機械学習手段は、教師データの事例について、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

、予め、何らかの手段により、学習結果記憶手段に記憶させる。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段のリストの優先順位の高い順に、抽出された解くべき問題の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解くべき問題の分類先（解）として推定する。

【０２２９】

また、機械学習手法として最大エントロピー法を使用する場合には、図示しない機械学習手段は、教師データの事例から解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となりうる分類の二項からなる確率分布を求めて、学習結果記憶手段に記憶する。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段の確率分布を利用して、抽出された解くべき問題の素性の集合についてその解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となりうる分類を特定し、その特定した分類をその解くべき問題の解と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

【０２３０】

また、機械学習手法としてサポートベクトルマシン法を使用する場合には、図示しない機械学習手段は、教師データの事例から解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したがって事例の素性の集合を次元とする空間上で、その事例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結果記憶手段に記憶する。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段の超平面を利用して、解くべき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もとづいて定まる分類を、その解くべき問題の解と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、分離平面からのその解くべき問題の事例への距離の大きさとする。

【０２３１】

また、上記実施の形態では、質問応答装置がスタンドアロンである場合について説明したが、質問応答装置は、スタンドアロンの装置であってもよく、サーバ・クライアントシステムにおけるサーバ装置であってもよい。後者の場合には、出力部や受付部は、通信回線を介して入力を受け付けたり、情報を出力したりすることになる。

【０２３２】

また、上記実施の形態において、各構成要素が実行する処理に関係する情報、例えば、各構成要素が受け付けたり、取得したり、選択したり、生成したり、送信したり、受信したりする情報や、各構成要素が処理で用いるしきい値や数式、アドレス等の情報等は、上記説明で明記していない場合であっても、図示しない記録媒体において、一時的に、あるいは長期にわたって保持されていてもよい。また、その図示しない記録媒体への情報の蓄積を、各構成要素、あるいは、図示しない蓄積部が行ってもよい。また、その図示しない記録媒体からの情報の読み出しを、各構成要素、あるいは、図示しない読み出し部が行ってもよい。

【０２３３】

また、上記実施の形態において、各処理または各機能は、単一の装置または単一のシステムによって集中処理されることによって実現されてもよく、あるいは、複数の装置または複数のシステムによって分散処理されることによって実現されてもよい。

【０２３４】

また、上記実施の形態において、質問応答装置に含まれる２以上の構成要素が通信デバイスや入力デバイス等を有する場合に、２以上の構成要素が物理的に単一のデバイスを有してもよく、あるいは、別々のデバイスを有してもよい。

【０２３５】

また、上記実施の形態において、各構成要素は専用のハードウェアにより構成されてもよく、あるいは、ソフトウェアにより実現可能な構成要素については、プログラムを実行

することによって実現されてもよい。例えば、ハードディスクや半導体メモリ等の記録媒体に記録されたソフトウェア・プログラムをＣＰＵ等のプログラム実行部が読み出して実行することによって、各構成要素が実現され得る。なお、上記実施の形態における質問応答装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、非ファクトイド（Ｎｏｎ－Ｆａｃｔｏｉｄ）型の質問を示す情報である質問情報を受け付ける質問情報受付部と、前記質問情報受付部が受け付けた質問情報に対して、当該質問情報の分類を示す情報であり、理由を尋ねる質問である理由質問が少なくとも一の分類として含まれる情報である複数の分類情報のいずれかを付与する分類部と、前記質問情報受付部が受け付けた質問情報から、用語を抽出する用語抽出部と、前記分類部が付与した分類情報に、対応情報記憶部で記憶される、分類を示す情報である分類情報と、前記用語抽出部が抽出した用語に追加する追加用語とを対応付けて有する情報である対応情報で対応付けられている追加用語と、前記用語抽出部が抽出した用語と、アクセス可能なコーパス記憶部で記憶されているコーパスと、前記分類部によって付与された分類情報に応じた式とを用いることによって、前記質問情報に対応する回答を示す情報である回答情報を前記コーパスから取得する回答情報取得部と、前記回答情報取得部が取得した回答情報を出力する回答情報出力部として機能させるためのものである。

【０２３６】

なお、上記プログラムにおいて、上記プログラムが実現する機能には、ハードウェアでしか実現できない機能は含まれない。例えば、情報を取得する取得部や、情報を出力する出力部などにおけるモデムやインターフェースカードなどのハードウェアでしか実現できない機能は、上記プログラムが実現する機能には少なくとも含まれない。

【０２３７】

また、このプログラムは、サーバなどからダウンロードされることによって実行されてもよく、所定の記録媒体（例えば、ＣＤ－ＲＯＭなどの光ディスクや磁気ディスク、半導体メモリなど）に記録されたプログラムが読み出されることによって実行されてもよい。

【０２３８】

また、このプログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【０２３９】

図１８は、上記プログラムを実行して、上記実施の形態による質問応答装置を実現するコンピュータの外観の一例を示す模式図である。上記実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムによって実現される。

【０２４０】

図１８において、コンピュータシステム１００は、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ Ｄｉｓｋ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）ドライブ１０５、ＦＤ（Ｆｌｅｘｉｂｌｅ Ｄｉｓｋ）ドライブ１０６を含むコンピュータ１０１と、キーボード１０２と、マウス１０３と、モニタ１０４とを備える。

【０２４１】

図１９は、コンピュータシステムを示す図である。図１９において、コンピュータ１０１は、ＣＤ－ＲＯＭドライブ１０５、ＦＤドライブ１０６に加えて、ＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）１１１と、ブートアッププログラム等のプログラムを記憶するためのＲＯＭ（Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）１１２と、ＣＰＵ１１１に接続され、アプリケーションプログラムの命令を一時的に記憶すると共に、一時記憶空間を提供するＲＡＭ（Ｒａｎｄｏｍ Ａｃｃｅｓｓ Ｍｅｍｏｒｙ）１１３と、アプリケーションプログラム、システムプログラム、及びデータを記憶するハードディスク１１４と、ＣＰＵ１１１、ＲＯＭ１１２等を相互に接続するバス１１５とを備える。なお、コンピュータ１０１は、ＬＡＮへの接続を提供する図示しないネットワークカードを含んでいてもよい。

【０２４２】

コンピュータシステム１００に、上記実施の形態による質問応答装置の機能を実行させ

るプログラムは、ＣＤ－ＲＯＭ１２１、またはＦＤ１２２に記憶されて、ＣＤ－ＲＯＭドライブ１０５、またはＦＤドライブ１０６に挿入され、ハードディスク１１４に転送されてもよい。これに代えて、そのプログラムは、図示しないネットワークを介してコンピュータ１０１に送信され、ハードディスク１１４に記憶されてもよい。プログラムは実行の際にＲＡＭ１１３にロードされる。なお、プログラムは、ＣＤ－ＲＯＭ１２１やＦＤ１２２、またはネットワークから直接、ロードされてもよい。

【０２４３】

プログラムは、コンピュータ１０１に、上記実施の形態による質問応答装置の機能を実行させるオペレーティングシステム（ＯＳ）、またはサードパーティプログラム等を必ずしも含んでいなくてもよい。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいてもよい。コンピュータシステム１００がどのように動作するのかについては周知であり、詳細な説明は省略する。

【０２４４】

また、本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【０２４５】

以上より、本発明による質問応答装置等によれば、非ファクトイド型の質問情報に対して適切に回答することができるという効果が得られ、質問情報に対応する回答情報を出力する装置等として有用である。

【図面の簡単な説明】

【０２４６】

【図１】本発明の実施の形態１による質問応答装置の構成を示すブロック図
【図２】同実施の形態による回答情報取得部の構成を示すブロック図
【図３】同実施の形態による質問応答装置の動作を示すフローチャート
【図４】同実施の形態による質問応答装置の動作を示すフローチャート
【図５】同実施の形態による質問応答装置の動作を示すフローチャート
【図６】同実施の形態による質問応答装置の動作を示すフローチャート
【図７】同実施の形態による質問応答装置の動作を示すフローチャート
【図８】同実施の形態による質問応答装置の動作を示すフローチャート
【図９】同実施の形態における対応情報の一例を示す図
【図１０】同実施の形態における表示の一例を示す図
【図１１】同実施の形態における第２の式のスコアの一例を示す図
【図１２】同実施の形態における取得された文書の一例を示す図
【図１３】同実施の形態における第１の式のスコアの一例を示す図
【図１４】同実施の形態における回答情報の表示の一例を示す図
【図１５】同実施の形態における分類対応情報の一例を示す図
【図１６】同実施の形態における実験結果の一例を示す図
【図１７】同実施の形態における機械学習について説明するための図
【図１８】同実施の形態におけるコンピュータシステムの外観一例を示す模式図
【図１９】同実施の形態におけるコンピュータシステムの構成の一例を示す図

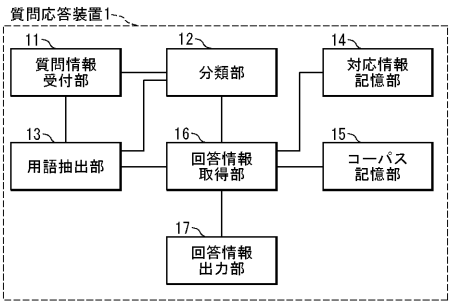
【符号の説明】

【０２４７】

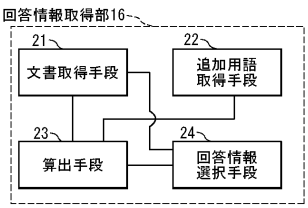
１　質問応答装置
１１　質問情報受付部
１２　分類部
１３　用語抽出部
１４　対応情報記憶部
１５　コーパス記憶部

１６　回答情報取得部
１７　回答情報出力部
２１　文書取得手段
２２　追加用語取得手段
２３　算出手段
２４　回答情報選択手段

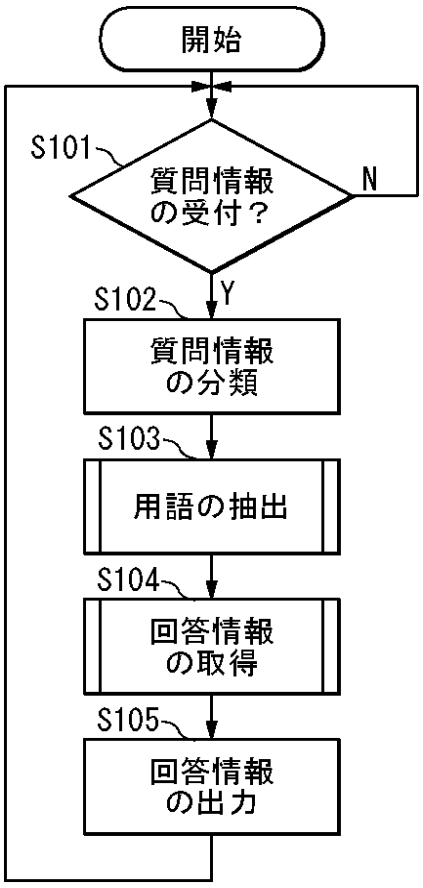
【図１】



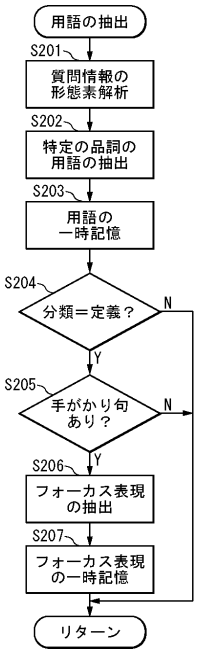
【図２】



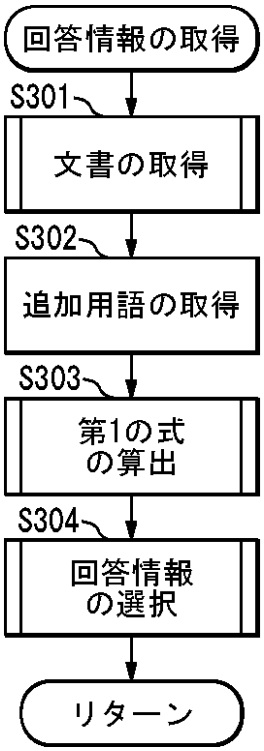
【図３】



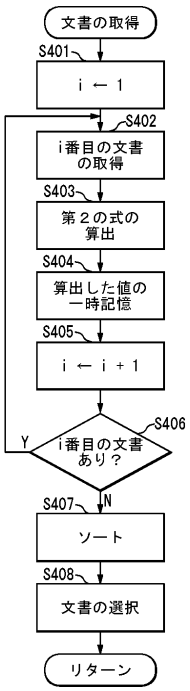
【図４】


用語の抽出
S201
質問情報の形態素解析
S202
特定の品詞の用語の抽出
S203
用語の一時記憶
S204
分類＝定義？
N
Y
S205
手がかり句あり？
N
Y
S206
フォーカス表現の抽出
S207
フォーカス表現の一時記憶
リターン

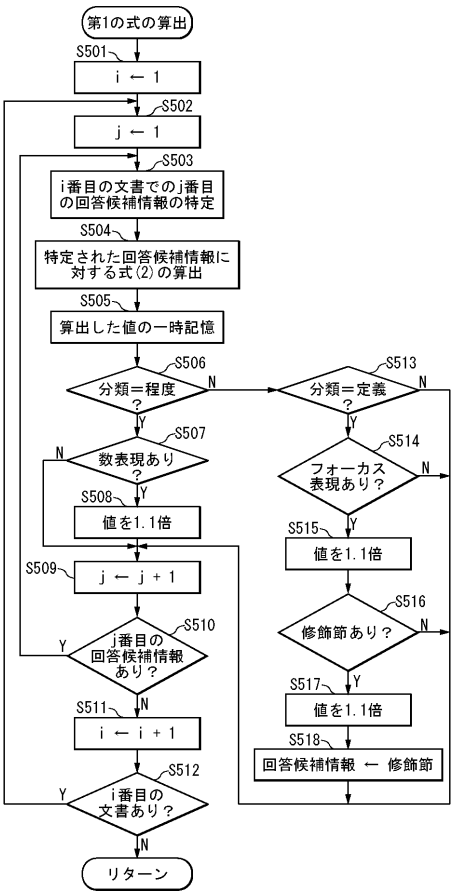
【図５】


回答情報の取得
S301
文書の取得
S302
追加用語の取得
S303
第1の式の算出
S304
回答情報の選択
リターン

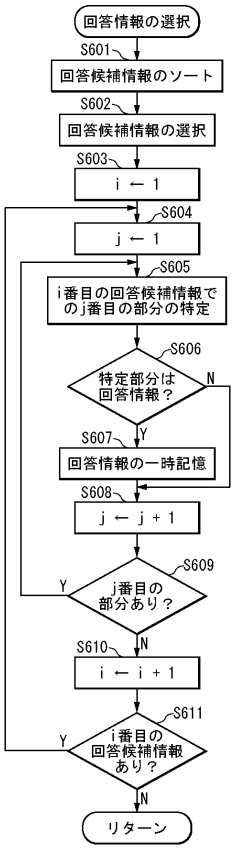
【図６】


文書の取得
S401
i ← 1
S402
i番目の文書の取得
S403
第２の式の算出
S404
算出した値の一時記憶
S405
i ← i + 1
S406
Y
i番目の文書あり？
S407
N
ソート
S408
文書の選択
リターン

【図７】


第1の式の算出
S501
i ← 1
S502
j ← 1
S503
i番目の文書でのj番目の回答候補情報の特定
S504
特定された回答候補情報に対する式(2)の算出
S505
算出した値の一時記憶
S506
分類＝程度？
N
Y
S507
数表現あり？
N
Y
S508
値を1.1倍
S509
j ← j + 1
S510
Y
j番目の回答候補情報あり？
S511
N
i ← i + 1
S512
Y
i番目の文書あり？
N
リターン
S513
分類＝定義？
N
Y
S514
フォーカス表現あり？
N
Y
S515
値を1.1倍
S516
修飾節あり？
N
Y
S517
値を1.1倍
S518
回答候補情報 ← 修飾節

【図８】



【図９】

対応情報

分類情報	追加用語
理由質問	理由
	原因
	なぜなら
方法質問	方法
	手順
	ことにより
経緯質問	経緯
	背景
	歴史

【図１０】

【図１１】

文書ID	式(1)のスコア
D001	2.3
D002	6.5
D003	4.2
D004	3.6
⋮	⋮

【図１２】

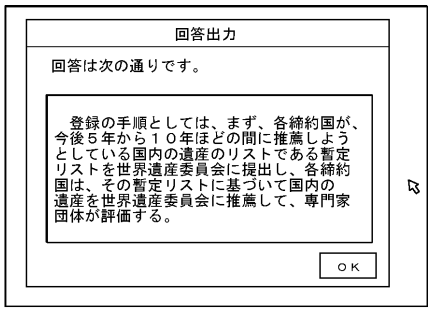
文書ID「D002」の文書

世界遺産とは、１９７２年の第１７回ユネスコ総会で採択された世界遺産条約（世界の文化遺産及び自然遺産の保護に関する条約）に基づいて、世界遺産リストに登録された文化遺産や自然遺産、文化的背景など、人類が共有すべきである普遍的な価値をもつものである。
世界遺産を保護する仕組みとして、……
：
：
登録の手順としては、まず、各締約国が、今後５年から１０年ほどの間に推薦しようとしている国内の遺産のリストである暫定リストを世界遺産委員会に提出し、各締約国は、その暫定リストに基づいて国内の遺産を世界遺産委員会に推薦して、専門家団体が評価する。そして、その評価結果に基づいて、世界遺産委員会が審査して、登録の可否を決定する。
２００６年現在の登録件数は、……
：
：

【図１３】

回答候補情報の識別情報		第1の式のスコア
文書ID	パラグラフ番号	
D002	1	2.6
	2	3.2
	3	9.7
	┆	┆
D003	1	3.6
	┆	┆
┆	┆	┆

【図１４】



【図１５】

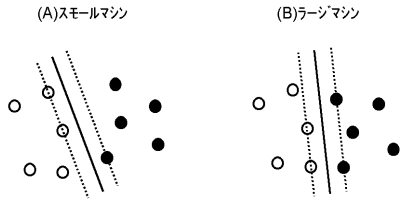
分類対応情報

分類情報	語句情報
定義質問	とは何
	どんな
	どういう
	どういった
	何もの
	┆
理由質問	なぜ
	なにゆえ
	どうして
	何が理由で
	どんな理由で
	┆
方法質問	どうすれば
	いかにして
	どうやって
	どのようにして
	┆
程度質問	どれくらい
	どの程度
	┆
変化質問	何がちがう
	どのように変わる
	どこが異なる
	┆
経緯質問	どのような経緯
	どのようないきさつ
	どのようななりゆき
	┆

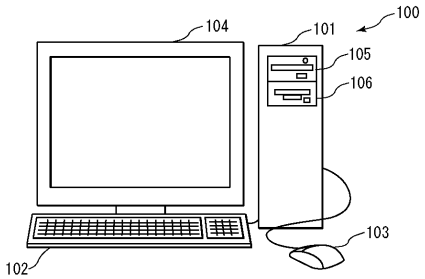
【図１６】

方法	正確	A	B	C	D
方法1	57	18	42	10	89
方法2	77	5	67	19	90

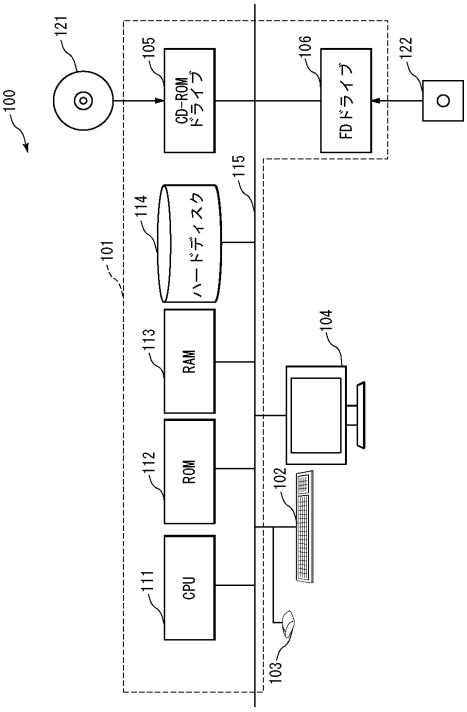
【図１７】



【図１８】



【図１９】



フロントページの続き

(56)参考文献 諸岡心，非Ｆａｃｔｏｉｄ型質問に対応した質問応答システム，言語処理学会第１３回年次大会発表論文集，日本，言語処理学会，２００７年 3月１９日，ｐｐ．９５８－９６１
渋沢潮，Ｗｈｙ型質問の回答文をＷｅｂページから抽出するシステムＲＥ：Ｗｈｙの試作，コンピュータソフトウェア，日本，日本ソフトウェア科学会，２００７年 7月２６日，ＶＯＬ．２４，ＮＯ．３，ｐｐ．２０－２８
石川開，質問同定を用いた自由文検索方式の提案 ～ コンタクトセンターＦＡＱ検索と携帯電話マニュアル音声検索 ～，マルチメディア，分散，協調とモバイル（ＤＩＣＯＭＯ２００７）シンポジウム論文集 情報処理学会シンポジウムシリーズ［ＣＤ－ＲＯＭ］，日本，社団法人情報処理学会 Information Processing Society of Japan，２００７年 6月２９日，Ｖｏｌ．２００７，Ｎｏ．１，ｐｐ．７９１－７９８
楊曄，語用情報に基づく『論語』の質問応答システムに関する研究，言語処理学会第１３回年次大会発表論文集，日本，言語処理学会，２００７年 3月１９日，ｐｐ．１０１０－１０１３

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０