

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3845727号

(P3845727)

(45) 発行日 平成18年11月15日(2006.11.15)

(24) 登録日 平成18年9月1日(2006.9.1)

(51) Int. Cl.	F I	
<b>G06F 17/30 (2006.01)</b>	G06F 17/30	210B
<b>G06F 17/28 (2006.01)</b>	G06F 17/30	170A
	G06F 17/30	180A
	G06F 17/30	330C
	G06F 17/30	340B
請求項の数 7 (全 26 頁) 最終頁に続く		

(21) 出願番号	特願2002-282795 (P2002-282795)	(73) 特許権者	301022471
(22) 出願日	平成14年9月27日(2002.9.27)		独立行政法人情報通信研究機構
(65) 公開番号	特開2004-118647 (P2004-118647A)		東京都小金井市貫井北町4-2-1
(43) 公開日	平成16年4月15日(2004.4.15)	(74) 代理人	100121511
審査請求日	平成14年9月27日(2002.9.27)		弁理士 小田 直
前置審査		(74) 代理人	100119161
			弁理士 重久 啓子
		(72) 発明者	村田 真樹
			東京都小金井市貫井北町4-2-1 独立行政法人通信総合研究所内
		審査官	和田 財太
			最終頁に続く

(54) 【発明の名称】 統計的検定を利用した質問応答システム

(57) 【特許請求の範囲】

## 【請求項1】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、

前記質問文を入力する質問文入力手段と、

入力された質問文の解が数量表現である場合に、統計的検定を用いて、前記電子化テキスト情報において出現する、前記質問文に含まれる、形態素解析と文パターンを用いた所定の規則に基づいて認定される特定の名詞と数量表現との組み合わせパターンから単位表現を抽出し、抽出された単位表現が前記質問文中の特定の名詞と共に出現する可能性を示す確率値を算出し、算出した前記確率値が閾値以上かによって、前記質問文の解としての数量表現の決定に用いる単位表現を決定し、決定した単位表現を含む数量表現を決定する手段と、

前記決定された数量表現を用いて前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする統計的検定を利用した質問応答システム。

## 【請求項2】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、

前記質問文を入力する質問文入力手段と、

前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、

10

20

形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の  
名詞を認定する手段と、

前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テ  
キスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせ  
パターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている  
1又は複数の単位表現を抽出する単位表現抽出手段と、

前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回  
数を算出し、算出した出現回数の情報に基づいて前記特定の名詞の一般的出現確率を算出  
する出現確率算出手段と、

前記算出した特定の名詞の一般的出現確率に基づいて、前記電子化テキスト情報におい  
て前記各単位表現が出現するパターンのうち前記特定の名詞が前記各単位表現と共に起して  
出現する回数の確率分布を算出する確率分布算出手段と、

前記算出した確率分布における前記特定の名詞が前記各単位表現と共に起して出現する回  
数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記  
数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、  
前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が  
予め設定した閾値以上の単位表現を除去する不要単位表現除去手段と、

前記検定確率が閾値未満の単位表現を含む数量表現を決定する手段と、

前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキ  
スト情報から前記質問文の解が記述された情報を取り出す手段と、

前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の  
解を抽出する解答抽出手段とを備える

ことを特徴とする統計的検定を利用した質問応答システム。

### 【請求項3】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文  
との照合によって解を生成して出力する質問応答システムであって、

前記質問文を入力する質問文入力手段と、

前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解  
が数量表現であるかを判断する数量表現判断手段と、

形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の  
名詞を認定する手段と、

前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テ  
キスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせ  
パターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている  
1又は複数の単位表現を抽出する単位表現抽出手段と、

前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回  
数を算出し、算出した出現回数の情報に基づいて前記各単位表現の一般的出現確率を算出  
する出現確率算出手段と、

前記各単位表現の一般的出現確率に基づいて、前記電子化テキスト情報において前記特  
定の名詞が出現するパターンのうち前記各単位表現が前記特定の名詞と共に起して出現する  
回数の確率分布を算出する確率分布算出手段と、

前記算出した確率分布における前記各単位表現が前記特定の名詞と共に起して出現する回  
数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおいて前記  
数量表現に付されている前記各単位表現と共に起して出現する前記数量表現の出現回数以上  
である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位  
表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を除去する不要単  
位表現除去手段と、

前記検定確率が閾値未満の単位表現を含む数量表現を決定する手段と、

前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキ  
スト情報から前記質問文の解が記述された情報を取り出す手段と、

前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備える

ことを特徴とする統計的検定を利用した質問応答システム。

【請求項 4】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、

前記質問文を入力する質問文入力手段と、

前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、

形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の  
10 名詞を認定する手段と、

前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている  
1 又は複数の単位表現を抽出する単位表現抽出手段と、

前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出し、算出した出現回数の情報に基づいて前記特定の名詞の一般的出現確率を算出する出現確率算出手段と、

前記算出した特定の名詞の一般的出現確率に基づいて、前記電子化テキスト情報において前記各単位表現が出現するパターンのうち前記特定の名詞が前記各単位表現と共に  
20 出現する回数の確率分布を算出する確率分布算出手段と、

前記算出した確率分布における前記特定の名詞が前記各単位表現と共に出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出手段が抽出した 1 又は複数の単位表現のうち、算出した前記検定確率が  
予め設定した閾値未満の単位表現を除去する不要単位表現除去手段と、

前記検定確率が閾値以上の単位表現を含む数量表現を決定する手段と、

前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、

前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の  
30 解を抽出する解答抽出手段とを備える

ことを特徴とする統計的検定を利用した質問応答システム。

【請求項 5】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、

前記質問文を入力する質問文入力手段と、

前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、

形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の  
40 名詞を認定する手段と、

前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている  
1 又は複数の単位表現を抽出する単位表現抽出手段と、

前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出し、算出した出現回数の情報に基づいて前記各単位表現の一般的出現確率を算出する出現確率算出手段と、

前記各単位表現の一般的出現確率に基づいて、前記電子化テキスト情報において前記特定の名詞が出現するパターンのうち前記各単位表現が前記特定の名詞と共に出現する  
50 回数の確率分布を算出する確率分布算出手段と、

前記算出した確率分布における前記各単位表現が前記特定の名詞と共起して出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおいて前記数量表現に付されている前記各単位表現と共起して出現する前記数量表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を除去する不要単位表現除去手段と、

前記検定確率が閾値以上の単位表現を含む数量表現を決定する手段と、

前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、

前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備える

10

ことを特徴とする統計的検定を利用した質問応答システム。

#### 【請求項6】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、

前記質問文を入力する質問文入力手段と、

前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、

形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、

20

前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、

前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出する手段と、

超幾何分布を用いて、超幾何分布における前記特定の名詞が前記各単位表現と共起して出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を除去する不要単位表現除去手段と、

30

前記検定確率が閾値未満の単位表現を含む数量表現を決定する手段と、

前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、

前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備える

ことを特徴とする統計的検定を利用した質問応答システム。

#### 【請求項7】

自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、

40

前記質問文を入力する質問文入力手段と、

前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、

形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、

前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、

前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回

50

数を算出する手段と、

超幾何分布を用いて、超幾何分布における前記特定の名詞が前記各単位表現と共に出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を除去する不要単位表現除去手段と、

前記検定確率が閾値以上の単位表現を含む数量表現を決定する手段と、

前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、

前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備える 10

ことを特徴とする統計的検定を利用した質問応答システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、コンピュータによる自然言語の情報処理システムに係わり、特に統計的検定を利用して不要な共起語を除去する質問応答システムに関するものである。

【0002】

質問応答システムとは、例えば「日本の首都はどこですか」、「2002年のワールドカップの優勝国はどこですか」といった質問を入力すると、大量の電子化テキストから「東京」、「ブラジル」といった解を的確に出力するシステムのことである。 20

【0003】

質問応答システムは、検索した記事から解を探す必要がある情報検索などとは異なり、解自体を的確に出力するため、ユーザーがより早く解の情報を得ることができる。また、質問応答システムは、解自体を自動で出力するため、他の自動の知識処理システムの内部での知識処理システムとして利用することも可能である。本発明は、このように有用な質問応答システムのうち、統計的検定を利用して不要な共起語を除去する質問応答システムに関する。

【0004】

【従来の技術】

質問応答システムの実行例として、「20世紀最初にトラはどのくらいいましたか」という質問文を与えることを考える。当該質問文の解は「10万頭」であるが、これを求めるために一般の質問応答システムでは「どのくらい」という表現から数量表現が解であろうと推測する。また、さらに「トラはどのくらい」また「どのくらいいました」といった表現から、「どのくらい」に相当する数量表現としては、「...頭」といった「頭」を単位表現とした数量表現であろうと推測する。

【0005】

この情報に基づいて、例えば、キーワード抽出を行ない、「20世紀最初」「トラ」をキーワードとして記事や文書を検索し、その検索によって得られた記事や文書から、数字表現+「頭」のパターンを取り出すことで、解の「10万頭」を取り出すことができる。 40

【0006】

このような単位表現を含んだ解を取り出す質問応答システムは従来から存在した。(非特許文献1参照。)

【0007】

また、与えられた文書集合を特徴付ける単語を選出する方法についても、従来から存在している(非特許文献2参照。)

【0008】

【非特許文献1】

佐々木裕，磯崎秀樹，平博順，平尾努，賀沢秀人，鈴木潤，国領弘治，前田英作，S A I Q A：大量文書に基づく質問応答システム，情報処理学会自然言語処理研究会2001 - 50

N L - 1 4 5 , 2 0 0 1

【非特許文献2】

久光徹，丹羽芳樹，組み合わせ的確率モデルに基づく特徴単語選択方法，情報処理学会自然言語処理研究会，140-12，2000

【0009】

【発明が解決しようとする課題】

しかし，上記非特許文献1に記載された従来技術は，単位表現の抽出に人手で記述した規則，もしくはテーブルを利用しているため，質問文によっては解の抽出に有用な単位表現かを的確に判断することができず，低い正解率しか得られない場合が生じ得る。

【0010】

また，非特許文献2に記載された従来技術は，文章のキーワードとしての特徴単語の選択を目的としており，言語の機能的な表現である単位表現の抽出を意図したものではなかった。

【0011】

本発明は，上記従来技術の問題点を解決し，質問文に対する解が数量表現の場合に，当該数量表現に付される単位表現として適切な単位表現か否かを的確に判断することが可能な質問応答方法および質問応答システムを提供することを目的とする。

【0012】

【課題を解決するための手段】

上記課題を解決するため，本発明では，質問文中の主たる名詞と単位表現との新聞コーパス等の電子化テキスト情報における共起頻度情報に基づいて，統計的検定を用いることにより解の抽出に有用な単位表現かを判断し，当該有用な単位表現のみを解の抽出に用い，不要と判断した単位表現を解の抽出に用いないようにする。主たる名詞とは，質問文中の質問の主たる対象をいい，例えば上記「20世紀最初にトラはどのくらいいましたか」という質問文においては「トラ」が主たる名詞である。

【0013】

ここで，例えば「20世紀最初にトラはどのくらいいましたか」という質問文の「どのくらい」に相当する数量表現は，「．．．頭」といった「頭」を単位表現とした数量表現であろうと推測する場合，「トラは」+数量表現」といったパターンを大量に取り出し，その数量表現に付されている単位表現を抽出することで単位表現「頭」を取り出したり，「数量表現+「いました」」といったパターンから同様に単位表現「頭」を取り出したりすることができる。

【0014】

しかし，このような方法だけでは，例えば，「トラは1992年に．．．」といった文からは「年」という単位表現が取り出されることになる。「年」も単位表現として解の抽出に使えることとしてしまうと，「1992年」といった表現を誤って解と出力する可能性がある。

【0015】

また，コーパス等における共起頻度情報を用いる方法も考えられる。「トラ」と共起して出現する単位表現の回数を数え，この回数の最も大きな単位表現のみを用いるのである。このようにすると，おそらく「頭」がもっとも出現頻度が高いので，「頭」を単位表現とすることになり，「年」を単位表現として抽出してしまうという問題は解消される。

【0016】

しかし，「20世紀最初にトラはどのくらいいましたか」という質問文の「どのくらい」に相当する数量表現での単位表現としては，「匹」という表現も考えられる。解の表現が「10万匹」となっていた場合，最大の頻度の「頭」だけを使うと，「10万匹」を解として取り出せなくなってしまう。

【0017】

そこで，本発明では，上記主たる名詞と単位表現とのコーパス等における共起頻度情報に基づき，統計的検定を用いて，前記数量表現に付されている各単位表現のうち前記主たる

10

20

30

40

50

名詞と共起して出現する可能性が低い単位表現を不要な共起語と判断して除去するとともに、有用と判断した単位表現のみを解の抽出に用いるようにする。

【0018】

具体的には、

『当該単位表現と共起して出現する「トラは」の出現確率 = コーパスにおける「トラは」の一般的出現確率』

という仮説を立てて、コーパス等における「トラは」と単位表現とが共起して出現する回数や「トラは」と単位表現のそれぞれの出現頻度に基づいて上記仮説の検定を行うことを通じて、「トラは」と「頭」、「匹」等の各単位表現が偶然共起したのか、必然的に共起しているものかの判断を行なう。検定結果に基づいて、「トラは」とある単位表現が偶然共起したものであると判断される場合、それは偶然共起しただけであり、不要な単位表現と判断する。また、「トラは」とある単位表現が必然的に共起したものであると判断される場合、それは必然的に共起したものであるから、関係が深い表現であろうと予想されるので、有用な単位表現と判断する。この方法により、個々の単位表現に対して、解の抽出に用いる単位表現として有用か不要かの判断を下すことができる。そして、有用と判断された単位表現のみを用いて質問応答システムの解の判断を行なうと、先のすべての問題が解消されることになる。

10

【0019】

即ち、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答方法であって、前記質問文を入力する質問文入力過程と、入力された質問文の解が数量表現である場合に、統計的検定を用いて、前記電子化テキスト情報において出現する前記質問文の主たる名詞と数量表現との組み合わせパターンから抽出される単位表現のうち前記質問文中の主たる名詞と共起して出現する可能性が低い単位表現を不要な共起語と判断し、前記質問文中の主たる名詞と共起して出現する可能性が高い単位表現を有用な共起語と判断し、前記有用な共起語と判断した単位表現を用いて前記質問文の解としての数量表現を推測する表現推測過程と、前記推測された数量表現を用いて前記質問文の解を抽出する解答抽出過程とを有することを特徴とする。

20

【0020】

また、本発明において、前記表現推測過程は、前記質問文情報に基づいて、質問文の解が数量表現であるかを判断する数量表現判断過程と、前記質問文情報から前記主たる名詞を認定する主名詞認定過程と、前記数量表現判断過程において質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出過程と、前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出し、算出した出現頻度情報に基づいて前記主たる名詞の一般的出現確率を算出する出現確率算出過程と、前記算出した主たる名詞の一般的出現確率に基づいて、前記電子化テキスト情報において前記各単位表現が出現するパターンのうち前記主たる名詞が前記各単位表現と共起して出現する回数の確率分布を算出する確率分布算出過程と、前記算出した確率分布における前記主たる名詞が前記各単位表現と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出過程において抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して除去し、前記検定確率が閾値未満の単位表現を有用な共起語と判断する不要単位表現除去過程と、前記有用と判断した単位表現を用いて前記質問文の解としての数量表現を推測する数量表現推測過程とを有することを特徴とする。

30

40

【0021】

また、本発明において、前記表現推測過程は、前記質問文情報に基づいて、質問文の解が

50

数量表現であるかを判断する数量表現判断過程と、前記質問文情報から前記主たる名詞を認定する主名詞認定過程と、前記数量表現判断過程において質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出過程と、前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出し、算出した出現頻度情報に基づいて前記各単位表現の一般的出現確率を算出する出現確率算出過程と、前記各単位表現の一般的出現確率に基づいて、前記電子化テキスト情報において前記主たる名詞が出現するパターンのうち前記各単位表現が前記主たる名詞と共起して出現する回数の確率分布を算出する確率分布算出過程と、前記算出した確率分布における前記各単位表現が前記主たる名詞と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおいて前記数量表現に付されている前記各単位表現と共起して出現する前記数量表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出過程において抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して除去し、前記検定確率が閾値未満の単位表現を有用な共起語と判断する不要単位表現除去過程と、前記有用と判断した単位表現を用いて前記質問文の解としての数量表現を推測する数量表現推測過程とを有することを特徴とする。

10

## 【0022】

20

また、本発明において、前記表現推測過程は、前記質問文情報に基づいて、質問文の解が数量表現であるかを判断する数量表現判断過程と、前記質問文情報から前記主たる名詞を認定する主名詞認定過程と、前記数量表現判断過程において質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出過程と、前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出し、算出した出現頻度情報に基づいて前記主たる名詞の一般的出現確率を算出する出現確率算出過程と、前記算出した主たる名詞の一般的出現確率に基づいて、前記電子化テキスト情報において前記各単位表現が出現するパターンのうち前記主たる名詞が前記各単位表現と共起して出現する回数の確率分布を算出する確率分布算出過程と、前記算出した確率分布における前記主たる名詞が前記各単位表現と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出過程において抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して除去し、前記検定確率が閾値以上の単位表現を有用な共起語と判断する不要単位表現除去過程と、前記有用と判断した単位表現を用いて前記質問文の解としての数量表現を推測する数量表現推測過程とを有することを特徴とする。

30

## 【0023】

40

また、本発明において、前記表現推測過程は、前記質問文情報に基づいて、質問文の解が数量表現であるかを判断する数量表現判断過程と、前記質問文情報から前記主たる名詞を認定する主名詞認定過程と、前記数量表現判断過程において質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出過程と、前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出し、算出した出現頻度情報に基づいて前記各単位表現の一般的出現確率を算出する出現確率算出過程と、前記各単位表現の一般的出現確率に基づいて、前記電子化テキスト情報において前記主たる名詞が出現するパターンのうち前記各単位表現が前記主たる名詞

50



と共起して出現する回数の確率分布を算出する確率分布算出過程と、前記算出した確率分布における前記各単位表現が前記主たる名詞と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおいて前記数量表現に付されている前記各単位表現と共起して出現する前記数量表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出過程において抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して除去し、前記検定確率が閾値以上の単位表現を有用な共起語と判断する不要単位表現除去過程と、前記有用と判断した単位表現を用いて前記質問文の解としての数量表現を推測する数量表現推測過程とを有することを特徴とする。

10

## 【0024】

また、本発明において、前記表現推測過程は、前記質問文情報に基づいて、質問文の解が数量表現であるかを判断する数量表現判断過程と、前記質問文情報から前記主たる名詞を認定する主名詞認定過程と、前記数量表現判断過程において質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出過程と、前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出する出現頻度抽出過程と、超幾何分布を用いて、超幾何分布における前記主たる名詞が前記各単位表現と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出過程において抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して除去し、前記検定確率が閾値未満の単位表現を有用な共起語と判断する不要単位表現除去過程と、前記有用と判断した単位表現を用いて前記質問文の解としての数量表現を推測する数量表現推測過程とを有することを特徴とする。

20

## 【0025】

また、本発明において、前記表現推測過程は、前記質問文情報に基づいて、質問文の解が数量表現であるかを判断する数量表現判断過程と、前記質問文情報から前記主たる名詞を認定する主名詞認定過程と、前記数量表現判断過程において質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出過程と、前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出する出現頻度抽出過程と、超幾何分布を用いて、超幾何分布における前記主たる名詞が前記各単位表現と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出過程において抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して除去し、前記検定確率が閾値以上の単位表現を有用な共起語と判断する不要単位表現除去過程と、前記有用と判断した単位表現を用いて前記質問文の解としての数量表現を推測する数量表現推測過程とを有することを特徴とする。

30

40

## 【0026】

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、入力された質問文の解が数量表現である場合に、統計的検定を用いて、前記電子化テキスト情報において出現する、前記質問文に含まれる、形態素解析と文パターンを用いた所定の規則に基づいて認定される特定の名詞と数量表

50

現との組み合わせパターンから単位表現を抽出し、抽出された単位表現が前記質問文中の特定の名詞と共起して出現する可能性を示す確率値を算出し、算出した前記確率値が閾値以上かによって、前記質問文の解としての数量表現の決定に用いる単位表現を決定し、決定した単位表現を含む数量表現を決定する手段と、前記決定された数量表現を用いて前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

**【0027】**

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出し、算出した出現回数の情報に基づいて前記特定の名詞の一般的出現確率を算出する出現確率算出手段と、前記算出した特定の名詞の一般的出現確率に基づいて、前記電子化テキスト情報において前記各単位表現が出現するパターンのうち前記特定の名詞が前記各単位表現と共起して出現する回数の確率分布を算出する確率分布算出手段と、前記算出した確率分布における前記特定の名詞が前記各単位表現と共起して出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を除去する不要単位表現除去手段と、前記検定確率が閾値未満の単位表現を含む数量表現を決定する手段と、前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

**【0028】**

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出し、算出した出現回数の情報に基づいて前記各単位表現の一般的出現確率を算出する出現確率算出手段と、前記各単位表現の一般的出現確率に基づいて、前記電子化テキスト情報において前記特定の名詞が出現するパターンのうち前記各単位表現が前記特定の名詞と共起して出現する回数の確率分布を算出する確率分布算出手段と、前記算出した確率分布における前記各単位表現が前記特定の名詞と共起して出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおいて前記数量表現に付されている前記各単位表現と共起して出現する前記数量表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を除去する不要単位表現除去手段と、前記検定確率が閾値未満の単位表現を含む数量表現を決定する手段と、前記質問文中からキーワード

10

20

30

40

50

を抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

【0029】

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出し、算出した出現回数の情報に基づいて前記特定の名詞の一般的出現確率を算出する出現確率算出手段と、前記算出した特定の名詞の一般的出現確率に基づいて、前記電子化テキスト情報において前記各単位表現が出現するパターンのうち前記特定の名詞が前記各単位表現と共に出現する回数の確率分布を算出する確率分布算出手段と、前記算出した確率分布における前記特定の名詞が前記各単位表現と共に出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を除去する不要単位表現除去手段と、前記検定確率が閾値以上の単位表現を含む数量表現を決定する手段と、前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

10

20

【0030】

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出し、算出した出現回数の情報に基づいて前記各単位表現の一般的出現確率を算出する出現確率算出手段と、前記各単位表現の一般的出現確率に基づいて、前記電子化テキスト情報において前記特定の名詞が出現するパターンのうち前記各単位表現が前記特定の名詞と共に出現する回数の確率分布を算出する確率分布算出手段と、前記算出した確率分布における前記各単位表現が前記特定の名詞と共に出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおいて前記数量表現に付されている前記各単位表現と共に出現する前記数量表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を除去する不要単位表現除去手段と、前記検定確率が閾値以上の単位表現を含む数量表現を決定する手段と、前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、前記決定された数量表現を用いて、前記質問文の解が記

30

40

50

述された情報から前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

【0031】

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出する手段と、超幾何分布を用いて、超幾何分布における前記特定の名詞が前記各単位表現と共に出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を除去する不要単位表現除去手段と、前記検定確率が閾値未満の単位表現を含む数量表現を決定する手段と、前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

10

20

【0032】

また、本発明は、自然言語による質問文を入力し、データベースに記憶された電子化テキスト情報中の文との照合によって解を生成して出力する質問応答システムであって、前記質問文を入力する質問文入力手段と、前記質問文に数量表現を問い合わせる単語が含まれるかどうかにより、前記質問文の解が数量表現であるかを判断する数量表現判断手段と、形態素解析と文パターンを用いた所定の規則に基づいて、前記質問文に含まれる特定の名詞を認定する手段と、前記数量表現判断手段が質問文の解が数量表現であると判断した場合に、前記電子化テキスト情報から前記認定された特定の名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出する単位表現抽出手段と、前記電子化テキスト情報における前記特定の名詞と前記抽出された各単位表現の出現回数を算出する手段と、超幾何分布を用いて、超幾何分布における前記特定の名詞が前記各単位表現と共に出現する回数が前記特定の名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以下である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値未満の単位表現を除去する不要単位表現除去手段と、前記検定確率が閾値以上の単位表現を含む数量表現を決定する手段と、前記質問文中からキーワードを抽出し、抽出したキーワードを用いて前記電子化テキスト情報から前記質問文の解が記述された情報を取り出す手段と、前記決定された数量表現を用いて、前記質問文の解が記述された情報から前記質問文の解を抽出する解答抽出手段とを備えることを特徴とする。

30

40

【0033】

また、本発明は、前記統計的検定を利用した質問応答方法をコンピュータに実行させるための統計的検定を利用した質問応答プログラムである。

【0034】

また、本発明は、前記統計的検定を利用した質問応答方法をコンピュータに実行させるための統計的検定を利用した質問応答プログラムを記録した記録媒体である。

【0035】

50

本発明を用いることにより、質問文に対する解が数量表現の場合に、当該数量表現に付される単位表現として適切な単位表現か否かを的確に判断することが可能な方法およびシステムを提供することが可能となる。

【0036】

【発明の実施の形態】

以下に、図を用いて、本発明の実施の形態を説明する。図1は本発明の統計的検定を利用した質問応答システムの構成の一例を示す図である。1は本発明の統計的検定を利用した質問応答システム、11は不要共起語除去システム、12は質問文情報が入力される質問文入力部、13はデータベース16からキーワード抽出や情報検索を行うキーワード抽出/情報検索部、14は質問文の解を抽出する解答抽出部、15は解答を出力する解答出力部、16は新聞コーパス等の電子化テキスト情報が記憶されたデータベース、17は質問文の解となり得そうな表現を推測する表現推測部、18はデータベース16から抽出した単位表現を記憶する単位表現データベースである。

10

【0037】

図2は、不要共起語除去システム11の構成図の一例である。20は質問文入力部12に入力された質問文情報に基づいて、質問文の解が数量表現であるかを判断する数量表現判断手段、21は質問文情報から主たる名詞を認定する主名詞認定手段、22はデータベース16に記憶された電子化テキスト情報から主名詞認定手段21が認定した主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現を抽出し、単位表現データベース18に記録する単位表現抽出手段、23は前記1又は複数の単位表現が前記主たる名詞と共起して出現する可能性が高いか否かを判断する検定手段、24は質問文の解としての数量表現を推測する数量表現推測手段である。

20

【0038】

また、230は前記電子化テキスト情報における前記主たる名詞と前記抽出された各単位表現の出現頻度を算出し、算出した出現頻度情報に基づいて前記主たる名詞の一般的出現確率を算出する出現確率算出手段、231は前記算出した主たる名詞の一般的出現確率に基づいて、前記電子化テキスト情報において前記各単位表現が出現するパターンのうち前記主たる名詞が前記各単位表現と共起して出現する回数の確率分布を算出する確率分布算出手段、232は確率分布算出手段231が算出した確率分布における前記主たる名詞が前記各単位表現と共起して出現する回数が前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンにおける前記数量表現に付されている各単位表現の出現回数以上である確率である検定確率を算出し、前記単位表現抽出手段が抽出した1又は複数の単位表現のうち、算出した前記検定確率が予め設定した閾値以上の単位表現を前記主たる名詞と共起して出現する可能性が低い不要な共起語と判断して単位表現データベース18から除去し、前記検定確率が閾値未満の単位表現を有用な単位表現と判断する不要単位表現除去手段である。

30

【0039】

以下に、図1、図2および図8を参照しつつ、図3および図4に基づいて本発明の第一の実施の形態を説明する。本発明の第一の実施の形態では、後述するように、各単位表現と共起する場合の主たる名詞のコーパスにおける出現確率 $p'$ が主たる名詞の一般的出現確率 $p$ と等しいという仮説を立てる。そして、一般的出現確率 $p$ に基づく主たる名詞がコーパスにおいて $n$ 回出現する各単位表現 $w_i$ と共起して出現する回数が、実際にコーパスにおいて各単位表現 $w_i$ と主たる名詞とが共起して出現する頻度である $x_i$ 回以上である確率(検定確率) $P$ の大きさに基づき上記仮説を右片側検定することを通じて $p' > p$ かを結論付け、 $p' > p$ と結論付けできる単位表現は有用な単位表現と判断し、 $p' > p$ と結論付けできない単位表現は不要な単位表現と判断する。

40

【0040】

図3は、本発明の統計的検定を利用した質問応答処理フローの一例を示す図である。まず、質問文が質問文入力部12に入力される(ステップS1)。例えば、「日本の国土面積

50

はどのくらいですか」という質問文が質問文入力部 1 2 に入力される。次に、入力された質問文情報が質問文入力部 1 2 から不要共起語除去システム 1 1 の表現推測部 1 7 に渡される(ステップ S 2)。次に、不要共起語除去システム 1 1 において、表現推測部 1 7 が、統計的検定を用いて、コーパスから抽出した単位表現のうち、解となり得そうな表現の推測に有用な単位表現かまたは不要な単位表現かを判断し、不要と判断した単位表現を除去する(ステップ S 3)。

【0041】

次に、表現推測部 1 7 が、有用と判断された単位表現を用いて、解となり得そうな表現を推測する(ステップ S 4)。具体的には、表現推測部 1 7 の数量表現推測手段 2 4 が、解となり得そうな数量表現を推測する。

10

【0042】

そして、キーワード抽出/情報検索部 1 3 が、表現推測部 1 7 から渡された質問文からキーワードを抽出する。そして、抽出したキーワードを用いてデータベースから解が記述してありそうな記事群を取り出し、取り出した記事群を解答抽出部 1 4 に渡す(ステップ S 5)。

【0043】

次に、解答抽出部 1 4 において、前記取り出された記事群から、表現推測部 1 7 で推測した表現に合致する表現を抽出し、抽出した表現を解答出力部 1 5 に渡す(ステップ S 6)。最後に、渡された表現を解答出力部 1 5 が解答として出力する(ステップ S 7)。

【0044】

ここで、本発明の第一の実施の形態における不要な単位表現の除去処理フローの詳細の一例を図 4 に示す。図 4 は、図 3 のステップ S 3 の詳細を示す図である。まず、数量表現判断手段 2 0 が、前記入力された質問文情報に基づいて、質問文の解が数量表現であるかを判断する(ステップ S 2 1)。例えば、質問文が「日本の国土面積はどのくらいですか」の場合のように「どのくらい」などの表現を含んでいた場合は、質問文の解が数量表現であると判断される。質問文の解が数量表現でないと判断する場合は前記ステップ S 5 へ進む。

20

【0045】

質問文の解が数量表現であると判断する場合は、数量表現判断手段 2 0 は、質問文が単位表現を有していないかを判断する(ステップ S 2 2)。質問文が単位表現を有していない場合には、主名詞認定手段 2 1 は、質問文情報から質問文における主たる名詞を認定する(ステップ S 2 3)。主たる名詞の認定は、「Xはどのくらい」のパターンから、形態素解析と文パターンを用いた規則に基づいてなされる。例えば上記質問文における主たる名詞は、「面積」であると認定される。質問文が単位表現を有している場合には、処理を終了する。

30

【0046】

次に、検定手段 2 3 の出現確率算出手段 2 3 0 は、前記コーパスにおける前記主たる名詞の一般的出現確率  $p$  を算出する(ステップ S 2 4)。かかる一般的出現確率  $p$  は、コーパスにおける前記主たる名詞の出現回数をコーパスの規模で割ったものである。毎日新聞などの大規模な新聞コーパスを用いると、例えば、上記「面積」のコーパスにおける出現回数

40

は 7, 9 3 3 回となる。コーパスの規模が、例えば 4 0 9, 5 0 2, 0 7 7 文字であるとすると、「面積」の一般的出現確率  $p$  は、

$$p = 7, 9 3 3 / 4 0 9, 5 0 2, 0 7 7$$

$$= 0 . 0 0 0 0 1 9 3 7 2 3 0 7 1 1 5 3 0 6 7 6$$

と算出される。

【0047】

次に、単位表現抽出手段 2 2 が、コーパスから前記主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている 1 又は複数の単位表現  $w_i$  を抽出し、単位表現データベース 1 8 に記録する(ステップ S 2 5)。本実施の形態では、「面積は」+数量表現のパターンをコー

50

パスから取り出し、数量表現に付されている単位表現  $w_i$  を取り出す。各単位表現  $w_i$  の取り出しは、形態素解析結果、また、単語辞書を用いて行なわれる。抽出された各単位表現  $w_i$  は、単位表現データベース 18 に記憶される。数量表現のパターンの取り出しは形態素解析と文パターンを用いた規則に基づいてなされる。

【0048】

ここで、コーパスから「面積は」+数量表現のパターンで取り出した数量表現から以下の単位表現  $w_i$  が例えば以下の頻度で抽出できる。

【0049】

平方メートル：41回、ヘクタール：27回、平方キロメートル：5回、倍：2回、分：2回、畳：1回、番：1回、平方センチメートル：1回

コロンの前が単位表現  $w_i$  で、数字が頻度である。「面積は」+数量表現のパターンの頻度は、上記の頻度の合計で80回である。

【0050】

次に、出現確率算出手段 230 が、上記抽出した各単位表現  $w_i$  のコーパスにおける出現頻度  $n$  を算出する(ステップ S26)。上記各単位表現  $w_i$  のコーパスでの出現頻度  $n$  は、例えば平方メートル：17,510回、ヘクタール：8,088回、平方キロメートル：247回、倍：41,686回、分：730,790回、畳：4,829回、番：124,233回、平方センチメートル：20回である。

【0051】

また、出現確率算出手段 230 は、実際にコーパスにおいて各単位表現  $w_i$  と主たる名詞とが共起して出現する頻度  $x_i$  を算出する(ステップ S27)。ここでの共起の定義は、例えば主たる名詞が「面積」の場合、「面積は」+数量表現のパターンにおける数量表現の単位表現が  $w_i$  であることとする。即ち、コーパスにおける各単位表現  $w_i$  の  $n$  個の出現パターンのうち、その数量表現と「面積」が「面積は」+数量表現"のパターンで共起して出現した頻度が  $x_i$  である。例えば、平方メートルの例だと、 $x_i = 41$  回となる。

【0052】

次に、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  = そのような条件のないときの「面積」の一般的出現確率  $p$  という仮説を立てて、二項分布を利用した検定を行う。即ち、まず、確率分布算出手段 231 が、前記算出した主たる名詞の一般的出現確率  $p$  に基づく前記コーパスにおいて前記主たる名詞と前記単位表現  $w_i$  とが共起して出現する回数の確率分布を算出する(ステップ S28)。個々の試行はすべて独立とし、ある単位表現  $w_i$  が  $n$  回出現しているときに、一般的出現確率  $p$  の主たる名詞が出現する回数の確率分布を求めるのである。かかる確率分布は、主たる名詞の出現回数を  $r$  ( $r = 0, 1, 2, 3 \dots n$ ) として、

$${}_n C_r p^r (1-p)^{n-r}$$

と算出され、図 8 に示されるような二項分布となる。

【0053】

次に、不要単位表現除去手段 232 は、算出した確率分布における前記主たる名詞と各単位表現  $w_i$  とが共起して出現する回数が  $x_i$  回以上である確率(検定確率)  $P$  を算出する(ステップ S29)。図 8 では斜線部分の確率が  $P$  である。例えば、「平方メートル」の  $n = 17510$  回の出現場面において、一般的出現確率  $p = \text{約} 0.00001937$  の「面積」が、 $x_i = 41$  回以上現れる検定確率  $P$  を求める。

【0054】

本発明の実施の形態では、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  とそのような条件のないときの「面積」の一般的出現確率  $p$  は本来異なる可能性があるのに同じであると仮定しており、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  に、そのような条件のないときの「面積」の一般的出現確率  $p$  を用いている。

【0055】

従って、検定確率  $P$  の値が十分に小さい場合は、 $p' = p$  という仮説は棄却され、 $p'$  と

10

20

30

40

50

$p$ とは異なると判断できる。また、本発明の実施の形態では、「 $x_i$ 回以上」のように、検定確率 $P$ を求める際に、確率分布の片側の領域のみを用いているので、片側検定になっている。さらに、用いている領域が「 $x_i$ 回以上」のように大きい場合の方の領域を用いる右片側検定であるため、 $P$ の値が十分に小さい場合、 $p' > p$ と結論付けできる。ただし、かかる判断は必ず正しいという意味ではなく、確率 $P$ だけは誤る可能性をもった判断である。

【0056】

不要単位表現除去手段232は、 $P$ の値が十分に小さく、「平方メートル」と共起する場合の「面積」の出現確率 $p'$ が、そのような条件のないときの「面積」の一般的出現確率 $p$ よりも大きいと結論付けできる場合は、この「平方メートル」と「面積」の共起は偶然ではなく必然的な共起であるとして「平方メートル」を有用な単位表現と判断する。

10

【0057】

逆に、検定確率 $P$ の値が十分に小さくなく、「平方メートル」と共起する場合の「面積」の出現確率 $p'$ が、そのような条件のないときの「面積」の一般的出現確率 $p$ よりも大きいと結論できない場合は、この「平方メートル」と「面積」の共起は必然的なものではなく偶然的な共起である可能性が高いとして「平方メートル」を不要な単位表現と判断し、以降の質問応答システムの処理では用いないようにする。

【0058】

即ち、不要単位表現除去手段232は、各単位表現 $w_i$ の検定確率 $P$ の値が予め設定した閾値未満かを判断し(ステップS30)、 $P$ の値が閾値未満の単位表現を有用な単位表現と判断し(ステップS32)、 $P$ の値が閾値以上の単位表現を不要な単位表現と判断して単位表現データベース18から除去し(ステップS31)、処理を終了する。

20

【0059】

実際に上記の例で $P$ の値を計算すると、平方メートル：0.00000000、ヘクタール：0.00000000、平方キロメートル：0.00000000、倍：0.19392510、分：0.99998923、畳：0.08930732、番：0.90988811、平方センチメートル：0.00038737となる。この計算は上述したように、二項分布の理論を使うことで計算することができる。ここでは、フリーソフトのMath-CDF-0.1のサブルーチンpbinomを用いて計算した。

【0060】

ここで、 $P$ の値が十分小さいことを意味する閾値として、例えば、0.1を用いると、

平方メートル：0.00000000

ヘクタール：0.00000000

平方キロメートル：0.00000000

畳：0.08930732

平方センチメートル：0.00038737

が有用な単位表現、すなわち妥当な共起語で、

倍：0.19392510

分：0.99998923

番：0.90988811

が不要な単位表現、すなわち妥当でない共起語であると判断できる。

30

40

【0061】

実際、平方メートル：0.00000000、ヘクタール：0.00000000、平方キロメートル：0.00000000、畳：0.08930732、平方センチメートル：0.00038737は、面積の単位表現として利用できるが、倍：0.19392510、分：0.99998923、番：0.90988811は、面積の単位表現としては不当な表現である。

【0062】

このようにして妥当な単位表現と判断した単位表現を用いて質問文の解としての数量表現の推測に利用するのである。

50



## 【0063】

本発明の第一の実施の形態においては、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  が「面積」の一般的出現確率  $p$  と等しいという仮説を立てて、 $P$  の値の大きさに基づいて右片側検定を行って、上記仮説が棄却できるか、即ち  $p' > p$  かを結論付け、 $p' > p$  と結論付けできる単位表現は有用な単位表現であり、 $p' > p$  と結論付けできない単位表現は不要な単位表現と判断していたが、本発明においては、以下のように左片側検定を行って  $p' < p$  かを結論付け、 $p' < p$  と結論付けできる単位表現は不要な単位表現であり、 $p' < p$  と結論付けできない単位表現は有用な単位表現と判断することもできる。

## 【0064】

本発明の第二の実施の形態においては、まず第一の実施の形態と同様に「面積」の表現の一般的出現確率  $p$  と、各単位表現の出現数  $n$  を求める。「平方メートル」の例では例えば  $n = 17, 510$  となる。

## 【0065】

この  $n$  個のパターンのうち、その数量表現と「面積」が「面積は」+数量表現" のパターンで共起して出現した頻度を  $x_i$  とする。例えば、平方メートルの例だと、 $x_i = 41$  となる。

## 【0066】

次に、個々の試行はすべて独立と仮定し、「平方メートル」の  $n$  回の出現場面において、一般的出現確率  $p$  の「面積」が、 $x_i$  回以下現れる確率（検定確率） $P$  を求める。言い換えると、「平方メートル」が  $n$  回出現していて、1 回に  $p$  の確率で出現する「面積」が、この  $n$  回の「平方メートル」とともに共起して出現する回数が  $x_i$  回以下である確率（検定確率） $P$  を求める。

## 【0067】

ここでは、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  とそのような条件のないときの「面積」の一般的出現確率  $p$  は本来異なる可能性があるのに同じであると仮定しており、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  に、そのような条件のないときの「面積」の一般的出現確率  $p$  を用いている。

## 【0068】

従って、検定確率  $P$  の値が十分に小さい場合は、 $p' = p$  という仮説は棄却され、 $p'$  と  $p$  とは異なると判断できる。また、本発明の実施の形態では、「 $x_i$  回以下」のように、検定確率  $P$  を求める際に、確率分布の片側の領域のみを用いているので、片側検定になっている。さらに、用いている領域が「 $x_i$  回以下」のように小さい場合の方の領域を用いる左片側検定であるので、 $P$  の値が十分に小さい場合、 $p' < p$  と結論付けできる。ただし、かかる判断は必ず正しいという意味ではなく、確率  $P$  だけは誤る可能性をもった判断である。

## 【0069】

不要単位表現除去手段 232 は、 $P$  の値が十分に小さく、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  が、そのような条件のないときの「面積」の一般的出現確率  $p$  よりも小さいと結論付けできる場合は、この「平方メートル」と「面積」の共起は偶然であるとして「平方メートル」を不要な単位表現と判断し、単位表現データベース 18 から除去する。

## 【0070】

逆に、検定確率  $P$  の値が十分に小さくなく、「平方メートル」と共起する場合の「面積」の出現確率  $p'$  が、そのような条件のないときの「面積」の一般的出現確率  $p$  よりも小さいと結論付けできない場合は、この「平方メートル」と「面積」の共起は偶然的なものではなく必然的な共起である可能性が高いとして「平方メートル」を有用な単位表現と判断し、当該有用と判断した単位表現を用いて数量表現推測手段 24 が質問文の解としての数量表現を推測する。

## 【0071】

10

20

30

40

50

以上整理すると、Pの値が小さいほど、不要な単位表現と判断し、Pの値が大きいほど、有用な単位表現と判断するということになる。

【0072】

実際に上記の例でPの値を計算すると、

平方メートル：1.00000000

ヘクタール：1.00000000

平方キロメートル：1.00000000

倍：0.95148679

分：0.00008198

畳：0.99588861

番：0.30698656

平方センチメートル：0.99999993

となる。

【0073】

この計算は二項分布の理論を使うことで計算することができる。ここでは、フリーソフトのMath-CDF-0.1のサブルーチンpbinomを用いて計算した。

【0074】

ここで、Pの値が十分小さいことを意味する閾値として、例えば、0.99を用いると、Pの値が0.99未満の単位表現は不要な単位表現であって不要な共起語であると判断し、Pの値が0.99以上の単位表現は主たる名詞「面積」と共起して出現する可能性が高く、有用な単位表現であって妥当な共起語であると判断する。即ち、

平方メートル：1.00000000

ヘクタール：1.00000000

平方キロメートル：1.00000000

畳：0.99588861

平方センチメートル：0.99999993

が妥当な共起語で、

倍：0.95148679

分：0.00008198

番：0.30698656

が妥当でない不要な共起語であると判断される。

【0075】

実際、平方メートル：1.00000000、ヘクタール：1.00000000、平方キロメートル：1.00000000、畳：0.99588861、平方センチメートル：0.99999993は面積の単位表現として利用できるが、倍：0.95148679、分：0.00008198、番：0.30698656は、面積の単位表現としては不当な表現である。

【0076】

このようにして求めた妥当な単位表現のみを用いて質問文の解としての数量表現の推測に利用するのである。

【0077】

図5は、上記本発明の第二の実施の形態における不要な単位表現の除去処理フローの詳細の一例を示す図であり、図3のステップS3の詳細であるステップS41乃至ステップS52を示したものである。なお、本発明の第二の実施の形態においては、図3におけるステップ3以外のステップは第一の実施の形態と同様である。

【0078】

図5に示すように、本発明の第二の実施の形態においては、ステップS49において、算出した確率分布における主たる名詞と各単位表現 $w_i$ とが共起して出現する回数が $x_i$ 回以下である確率P(検定確率)を算出することと、ステップS52においてPの値<閾値の場合には不要な単位表現と判断して単位表現データベースから除去することと、ステッ

10

20

30

40

50

プ S 5 1 において P の値 < 閾値でない場合には有用な単位表現と判断する点において図 4 に示す本発明の第一の実施の形態と異なり、図 5 における他のステップは本発明の第一の実施の形態と同様である。

【 0 0 7 9 】

本発明は、その趣旨に基づき、以下のように種々の変形が可能である。

【 0 0 8 0 】

上記本発明の第一の実施の形態および第二の実施の形態では、一般的出現確率を  $p$  として、コーパスでの主たる名詞としての「面積」の出現確率を用いたが、本発明では、コーパスにおける一般的出現確率  $p$  としては、単位表現  $w_i$  の出現確率を用いて、コーパスにおいて主たる名詞「面積」が出現した個数を  $n$  としても同様の検定が行なえる。

10

【 0 0 8 1 】

即ち、「面積」と共起する場合の「平方メートル」の出現確率  $p'$  が「平方メートル」の一般的出現確率  $p$  と等しいという仮説を立てて左片側検定を行い、「平方メートル」が上記  $n$  回出現する主たる名詞「面積」と共起して出現する回数が  $x_i$  回（例えば 4 1 回）以下である検定確率  $P$  の値の大きさが閾値（例えば 0.99）未満である場合は  $p' < p$  であると結論付けて「平方メートル」は不要な単位表現と判断し、検定確率  $P$  の値の大きさが閾値以上であり、 $p' < p$  と結論付けできない場合は有用な単位表現と判断することもできる。

【 0 0 8 2 】

例えば、「平方メートル」の場合だと、

$p = 17510 / 409502077$ 、 $n = 7933$  であり、

$P = 1.00000000 > 0.99$

となつて、 $p' < p$  と判断できないため、「平方メートル」は有用な単位表現であると判断できる。

20

【 0 0 8 3 】

かかる左片側検定を行う本発明の第三の実施の形態の不要な単位表現の除去処理フローの詳細の一例を図 6 に示す。図 6 は、図 3 のステップ S 3 の詳細であるステップ S 8 1 乃至ステップ S 9 2 を示したものである。図 6 に示すように、本発明の第三の実施の形態においては、ステップ S 8 4 においてコーパスにおける各単位表現  $w_i$  の一般的出現確率  $p$  を算出していることと、ステップ S 8 6 において、主たる名詞のコーパスにおける出現頻度  $n$  を算出している点において図 5 に示す本発明の第二の実施の形態と異なる。

30

【 0 0 8 4 】

もちろん、本発明においては、「面積」と共起する場合の「平方メートル」の出現確率  $p'$  が「平方メートル」の一般的出現確率  $p$  と等しいという仮説を立てて右片側検定を行い、「平方メートル」が上記  $n$  回出現する「面積」と共起して出現する回数が  $x_i$  回以上である検定確率  $P$  の値の大きさが閾値未満である場合は  $p' > p$  と結論付けて有用な単位表現と判断し、検定確率  $P$  の値の大きさが閾値以上であり、 $p' > p$  と結論付けできない場合は不要な単位表現と判断することもできる。

【 0 0 8 5 】

かかる右片側検定を行う本発明の第四の実施の形態の不要な単位表現の除去処理フローの詳細の一例を図 7 に示す。図 7 は、図 3 のステップ S 3 の詳細であるステップ S 6 1 乃至ステップ S 7 2 を示したものである。図 7 に示すように、本発明の第四の実施の形態においては、ステップ S 6 4 においてコーパスにおける各単位表現  $w_i$  の一般的出現確率  $p$  を算出していることと、ステップ S 6 6 において、主たる名詞のコーパスにおける出現頻度  $n$  を算出している点において図 4 に示す本発明の第一の実施の形態と異なる。

40

【 0 0 8 6 】

また、本発明においては、以下に示す超幾何分布を用いた検定を行うこともできる。

【 0 0 8 7 】

超幾何分布とは、

$$h g ( N , k , n , m ) = C ( K , m ) \times C ( N - K , n - m ) / C ( N , n )$$

50

の形で表せる分布である。

【0088】

ただし、 $C$  は「組み合わせ」を意味する記号  $C(A, B) = A! / B! / (B - A)!$  であり、また、 $hg(N, k, n, m)$  は、「 $N$ 個の玉の中に $K$ 個の赤い玉があるとき、任意に取り出した $n$ 個の玉の中に赤い玉がちょうど $m$ 個含まれる確率」である。

【0089】

ここで、 $hgs(N, K, n, k) = hg(N, K, n, m)$

とすると、 $hgs(N, K, n, k)$  は、「 $N$ 個の玉の中に $K$ 個の赤い玉があるとき、任意に取り出した $n$ 個の玉の中に赤い玉が $k$ 個以上含まれる確率」となる。なお、 $hg(N, K, n, m)$  は、 $m \geq k$  である $m$ についての $hg(N, k, n, m)$ の合計を意味する。

10

【0090】

ここで、 $N, K, n, k$  を以下のように解釈する。

【0091】

$N$  : コーパスの大きさ

$K$  : 「面積」の出現頻度

$n$  : ある単位表現の出現頻度

$k$  : 「面積」と、ある単位表現の共起回数

上記解釈によると、 $hgs(N, K, n, k)$  は、「 $N$ の大きさのコーパスの中に「面積」という表現が $K$ 個あるときに、ある着目している単位表現を $n$ 個取り出し、その $n$ 個の単位表現のうち、 $k$ 個以上のもので、「面積」とその単位表現が共起する事象の起る確率」を意味することとなる。

20

【0092】

これは、前記本発明の実施の形態において、 $p = K / N$  を「面積」の一般的出現確率としていたところを、 $K / N$  の形にまとめずに $K$ と $N$ にわけたまま扱っていることに相当する。超幾何分布を用いる方法では、 $p = K / N$  とせずに $K$ と $N$ とにわけている分、仮定が少なく、近似の少ない手法で精度は高くなるものである。

【0093】

この方法では、 $hgs(N, K, n, k)$  が小さいほど、「面積」と単位表現の共起が妥当なものと判断でき、 $hgs(N, K, n, k)$  が大きいほど、「面積」と単位表現の共起は偶然のもので、妥当なものではないと判断できる。

30

【0094】

従って、本発明においては、不要単位表現除去手段232は、 $hgs(N, K, n, k)$  がある閾値より小さい単位表現を有用な単位表現と判断して質問文の解としての数量表現の推定に用いるようにし、 $hgs(N, K, n, k)$  がある閾値以上の単位表現を不要な単位表現と判断して除去する。

【0095】

本発明の第五の実施の形態では、上述した超幾何分布を用いて右片側検定を行う。図9は、本発明の第五の実施の形態における不要な単位表現の除去処理フローの詳細の一例を示す図であり、図3のステップS3の詳細であるステップS101乃至ステップS111を示したものである。なお、図3におけるその他のステップは本発明の第一の実施の形態と同様である。

40

【0096】

図9に示すように、本発明の第五の実施の形態においては、質問文の解が数量表現であるかを判断し(ステップS101)、解が数量表現でない場合は図3のステップS5に移行し、解が数量表現の場合は質問文が単位表現を有していないかを判断する(ステップS102)。

【0097】

質問文が単位表現を有している場合には処理を終了して図3のステップS4に移行し、質問文が単位表現を有していない場合には、質問文情報から質問文における主たる名詞を認

50

定する（ステップS103）。

【0098】

次に、コーパスにおける主たる名詞の出現頻度 $K$ を算出する（ステップS104）。そして、コーパスから主たる名詞と数量表現とが係り受け関係である組み合わせパターンを抽出し、抽出された前記組み合わせパターンから前記数量表現に付されている1又は複数の単位表現 $w_i$ を抽出し、単位表現データベース18に記録する（ステップS105）。そして、抽出した各単位表現 $w_i$ のコーパスにおける出現頻度 $n$ を算出する（ステップS106）。

【0099】

次に、各単位表現 $w_i$ と主たる名詞とが共起して出現する回数 $k$ を算出する（ステップS107）。次に、超幾何分布における主たる名詞と各単位表現 $w_i$ とが共起して出現する回数が $k$ 回以上である検定確率 $P = h g s ( N , K , n , k )$ を算出する（ステップS108）。なお、 $N$ はコーパスの大きさである。

【0100】

そして、 $P$ の値<閾値であるかを判断し（ステップS109）、 $P$ の値<閾値である場合には有用な単位表現と判断し（ステップS111）、 $P$ の値<閾値でない場合は、不要な単位表現と判断して単位表現データベースから除去する（ステップS110）。

【0101】

また、本発明においては、超幾何分布を用いて左片側検定を行うこともできる。超幾何分布を用いて左片側検定を行う本発明の第六の実施の形態における不要単位表現の除去処理フローの詳細の一例を図10に示す。図10は、図3のステップS3の詳細であるステップS121乃至ステップS131を示したものである。本発明の第六の実施の形態においては、図10のステップS128において超幾何分布における主たる名詞と各単位表現 $w_i$ とが共起して出現する回数が $k$ 回以下である検定確率 $P = h g s ( N , K , n , k )$ を算出することと、ステップS130において、 $P$ の値<閾値でない場合には有用な単位表現と判断することと、ステップS131において、 $P$ の値<閾値である場合は、不要な単位表現と判断して単位表現データベースから除去する点で図9に示す本発明の第五の実施の形態と異なる。

【0102】

なお、本発明の第五の実施の形態または本発明の第六の実施の形態においては、 $K$  = 主たる名詞の出現頻度、 $n$  = 各単位表現の出現頻度としたが、この $K$ と $n$ は交換可能で、 $K$  = 各単位表現の出現頻度、 $n$  = 主たる名詞の出現回数としてもよい。

【0103】

以上説明した全ての本発明の実施の形態では、ある種の比率の検定を行なっていることに相当する。おおよそ、主たる名詞と単位表現の一般的出現確率から予想される主たる名詞と単位表現の共起回数よりも大きい回数もしくは共起回数以上の回数で主たる名詞と単位表現の共起が出現しているかどうかを検定するのである。

【0104】

また、本発明は、AICやZ-scoreなどの比率の検定を行なうことができる統計的検定法でも実現できる。

【0105】

なお、本発明は、前記従来技術における人手で記述した規則、もしくは、テーブルの作成の補助にも用いることができる。

【0106】

【発明の効果】

本発明を用いることにより、質問文に対する解が数量表現の場合に、当該数量表現に付される単位表現として適切な単位表現か否かを的確に判断することが可能な方法およびシステムを提供することが可能となる。

【図面の簡単な説明】

【図1】統計的検定を利用した質問応答システムの構成の一例を示す図である。

10

20

30

40

50

- 【図2】不要共起語除去システムの構成図の一例である。
- 【図3】統計的検定を利用した質問応答処理フローの一例を示す図である。
- 【図4】不要な単位表現の除去処理フローの詳細の一例を示す図である。
- 【図5】不要な単位表現の除去処理フローの詳細の一例を示す図である。
- 【図6】不要な単位表現の除去処理フローの詳細の一例を示す図である。
- 【図7】不要な単位表現の除去処理フローの詳細の一例を示す図である。
- 【図8】二項分布を示す図である。
- 【図9】不要な単位表現の除去処理フローの詳細の一例を示す図である。
- 【図10】不要な単位表現の除去処理フローの詳細の一例を示す図である。

【符号の説明】

1 統計的検定を利用した質問応答システム

1 1 不要共起語除去システム

1 2 質問文入力部

1 3 キーワード抽出 / 情報検索部

1 4 解答抽出部

1 5 解答出力部

1 6 データベース

1 7 表現推測部

1 8 単位表現データベース

2 0 数量表現判断手段

2 1 主名詞認定手段

2 2 単位表現抽出手段

2 3 検定手段

2 4 数量表現推測手段

2 3 0 出現確率算出手段

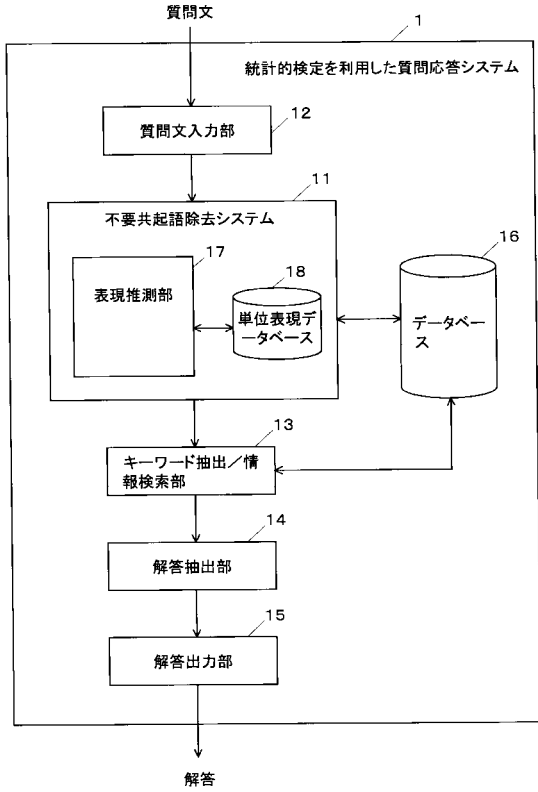
2 3 1 確率分布算出手段

2 3 2 不要単位表現除去手段

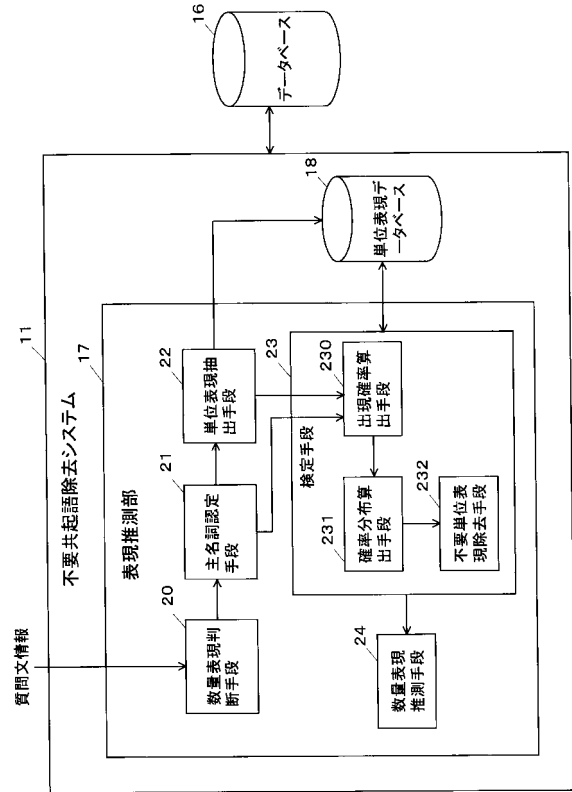
10

20

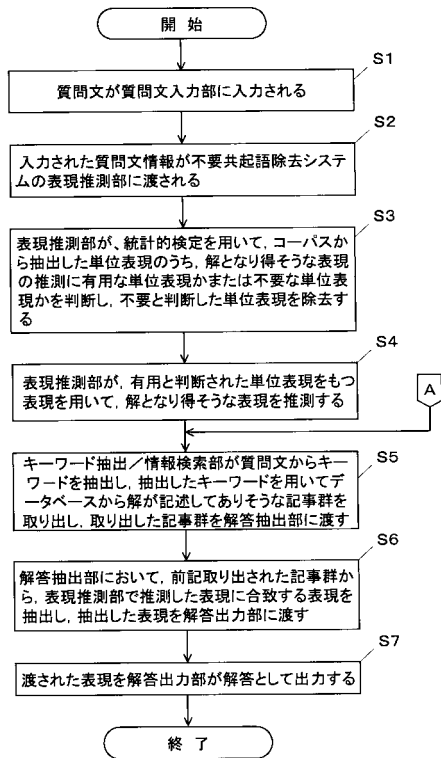
【 図 1 】



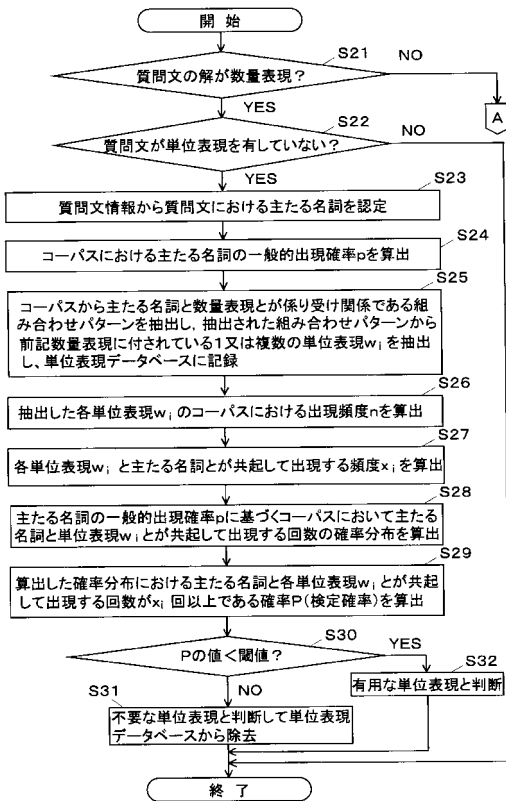
【 図 2 】



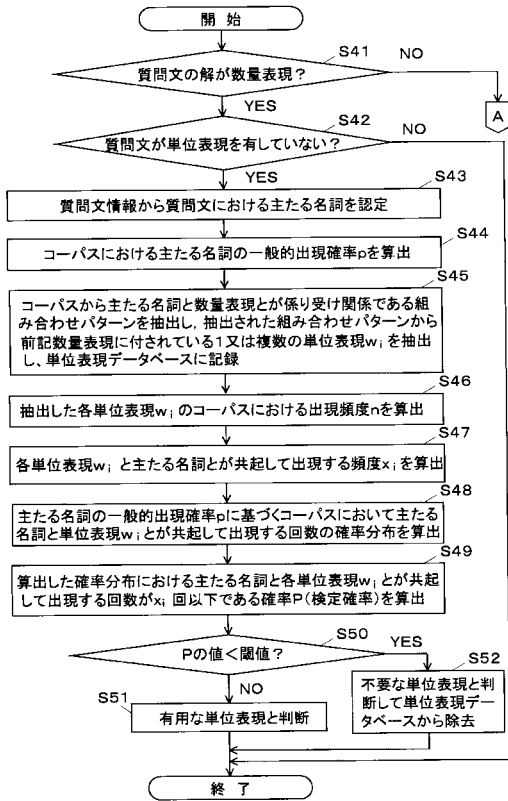
【 図 3 】



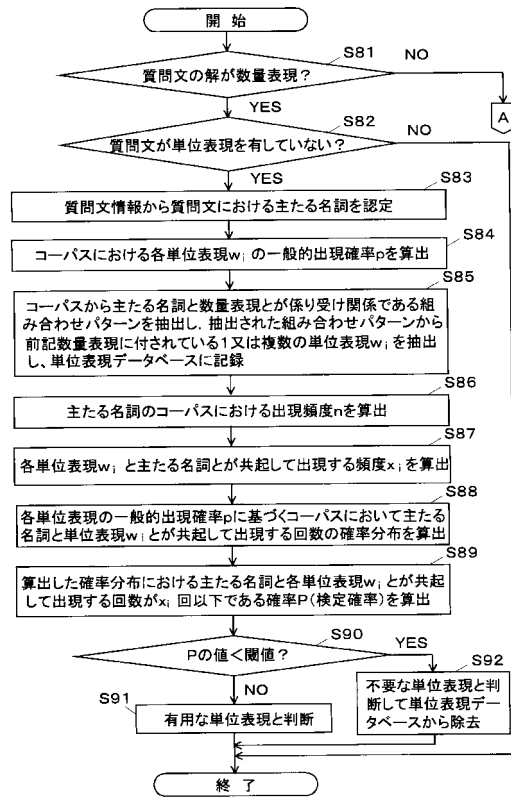
【 図 4 】



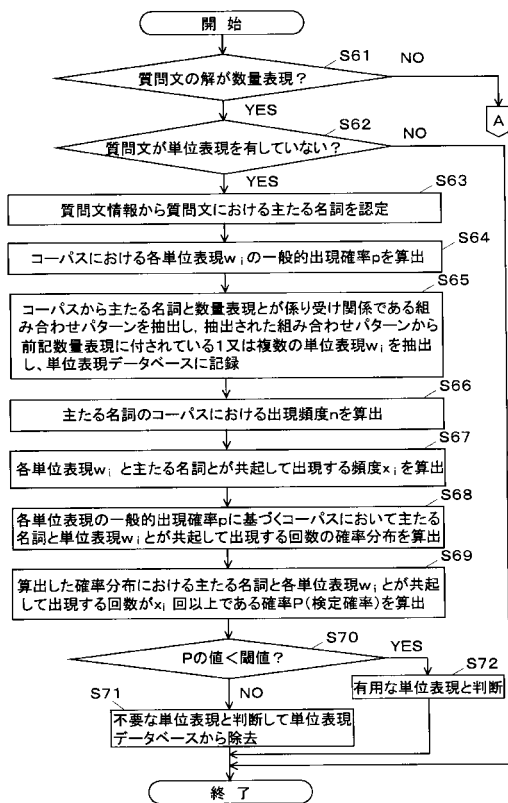
【図5】



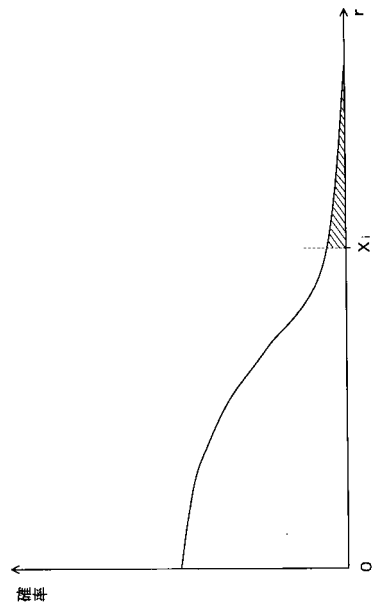
【図6】



【図7】

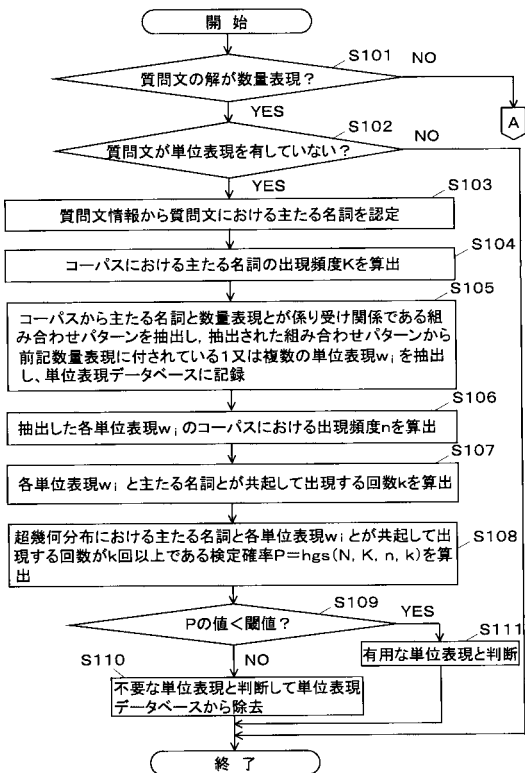


【図8】

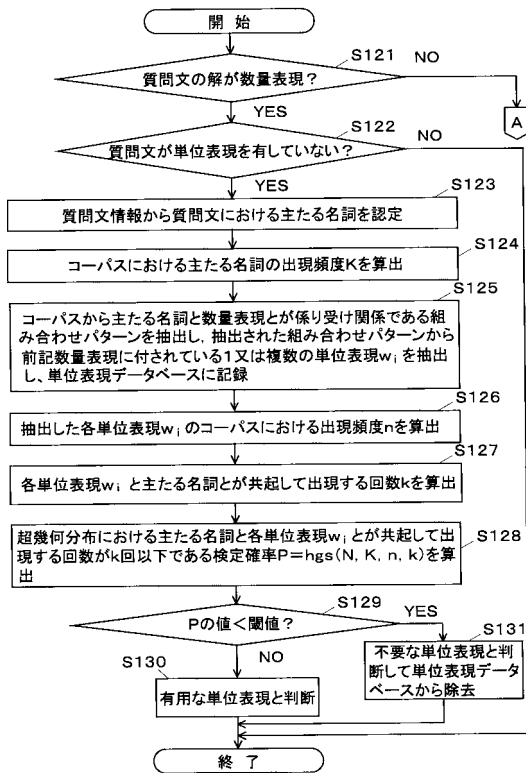




【 図 9 】



【 図 10 】



## フロントページの続き

(51) Int.Cl.

F I

G 0 6 F 17/28

X

- (56) 参考文献 久光徹・丹羽芳樹，組み合わせの確率モデルに基づく特徴単語選択方法，情報処理学会研究会報告2000-NL-140-12，日本，社団法人情報処理学会，2000年11月21日，Vol.2000，No.107，p.85-p.90  
佐々木裕・磯崎秀樹・平博順・平尾努・賀沢秀人・鈴木潤・国領弘治・前田英作，SAIQA：大量文書に基づく質問応答システム，情報処理学会研究報告2001-NL-145-12，日本，社団法人情報処理学会，2001年9月10日，Vol.2001，No.86，p.77-p.82

(58) 調査した分野(Int.Cl.，DB名)

G06F 17/21-17/30