(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-118647
(P2004-118647A)
(43)公開日　平成16年4月15日(2004.4.15)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	210B	5B075
	G06F 17/30	170A	
	G06F 17/30	180A	
	G06F 17/30	330C	
	G06F 17/30	340B	

審査請求 有　請求項の数 16　OL　（全 27 頁）

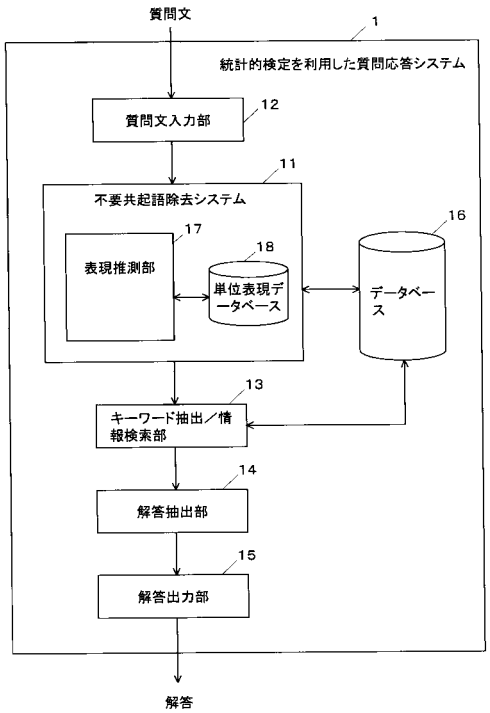
(21)出願番号　特願2002-282795(P2002-282795)
(22)出願日　平成14年9月27日(2002.9.27)

(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(74)代理人　100121511
弁理士　小田　直
(74)代理人　100097836
弁理士　福井　國敏
(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

Fターム(参考)　5B075 ND03　NK32　NK33　PP24　PR04
QP01　QP03

(54)【発明の名称】統計的検定を利用した質問応答方法，質問応答システム，質問応答プログラムおよび質問応答プログラムを記録した記録媒体

(57)【要約】
【課題】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。
【解決手段】入力された質問文情報とデータベース１６に記憶された電子化テキスト情報とに基づいて，不要共起語削除システム１１の表現推測部１７において，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単位表現と判断して除去する。
【選択図】　図１



【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
ことを特徴とする統計的検定を利用した質問応答方法。
【請求項２】
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
ことを特徴とする請求項１記載の統計的検定を利用した質問応答方法。
【請求項３】
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
ことを特徴とする請求項１記載の統計的検定を利用した質問応答方法。
【請求項４】
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
ことを特徴とする請求項１記載の統計的検定を利用した質問応答方法。
【請求項５】
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
ことを特徴とする請求項１記載の統計的検定を利用した質問応答方法。
【請求項６】
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
ことを特徴とする請求項１記載の統計的検定を利用した質問応答方法。
【請求項７】
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
ことを特徴とする請求項１記載の統計的検定を利用した質問応答方法。
【請求項８】
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前記質問文の解としての数量表現の推定に用いる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手段と，
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９】
前記表現推測手段は，
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手段と，
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手段と，
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手段と，
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手段と，
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手段と，
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手段とを備える
ことを特徴とする請求項８記載の統計的検定を利用した質問応答システム。
【請求項１０】
前記表現推測手段は，
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手段と，
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手段と，
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手段と，
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手段とを備える
ことを特徴とする請求項８記載の統計的検定を利用した質問応答システム。
【請求項１１】
前記表現推測手段は，
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手段と，
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手段と，
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手段と，
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手段と，
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手段と，
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手段とを備える
ことを特徴とする請求項８記載の統計的検定を利用した質問応答システム。
【請求項１２】
前記表現推測手段は，
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手段と，
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手段と，
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手段と，
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手段とを備える
ことを特徴とする請求項８記載の統計的検定を利用した質問応答システム。
【請求項１３】
前記表現推測手段は，
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手段と，
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手段と，
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手段と，
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手段と，
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手段とを備える
ことを特徴とする請求項８記載の統計的検定を利用した質問応答システム。
【請求項１４】
前記表現推測手段は，
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手段と，
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手段と，
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手段と，

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手段と，
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手段とを備える
ことを特徴とする請求項８記載の統計的検定を利用した質問応答システム。

【請求項１５】
請求項１から請求項７までのいずれか１項に記載の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。

【請求項１６】
請求項１から請求項７までのいずれか１項に記載の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。

【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。

【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。

【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，ＳＡＩＱＡ：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－ＮＬ－１４５，２００１
【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ものか，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。

【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。

【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。

【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。

【００２２】

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。

【００２３】

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。

【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。

【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。

【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前記質問文の解としての数量表現の推定に用いる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手段と，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手段とを備えることを特徴とする。

【００２７】

また，本発明において，前記表現推測手段は，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手段と，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手段と，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手段と，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手段と，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手段と，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手段とを備えることを特徴とする。

【００２８】

また，本発明において，前記表現推測手段は，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手段と，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手段と，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手段と，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手段とを備えることを特徴とする。

【００２９】
また，本発明において，前記表現推測手段は，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手段と，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手段と，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手段と，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手段と，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手段と，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手段とを備えることを特徴とする。

【００３０】
また，本発明において，前記表現推測手段は，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手段と，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手段と，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手段と，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手段とを備えることを特徴とする。

【００３１】
また，本発明において，前記表現推測手段は，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手段と，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手段と，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手段と，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手段と，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手段とを備えることを特徴とする。

【００３２】

また，本発明において，前記表現推測手段は，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手段と，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手段と，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手段と，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手段と，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手段とを備えることを特徴とする。

【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】

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。

【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。

【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。

【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。

【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，「日本の国土面積はどのくらいですか」という質問文が質問文入力部１２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１２から不要共起語除去システム１１の表現推測部１７に渡される（ステップＳ２）。次に，不要共起語除去システム１１において，表現推測部１７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Ｓ３）。

【００４１】

次に，表現推測部１７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Ｓ４）。具体的には，表現推測部１７の数量表現推測手段２４が，解となり得そうな数量表現を推測する。

【００４２】

そして，キーワード抽出／情報検索部１３が，表現推測部１７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１４に渡す（ステップＳ５）。

【００４３】

次に，解答抽出部１４において，前記取り出された記事群から，表現推測部１７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１５に渡す（ステップＳ６）。最後に，渡された表現を解答出力部１５が解答として出力する（ステップＳ７）。

【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Ｓ３の詳細を示す図である。まず，数量表現判断手段２０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Ｓ２１）。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Ｓ５へ進む。
【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２０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Ｓ２２）。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２１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Ｓ２３）。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。
【００４６】
次に，検定手段２３の出現確率算出手段２３０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Ｓ２４）。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は７，９３３回となる。コーパスの規模が，例えば４０９，５０２，０７７文字であるとすると，「面積」の一般的出現確率ｐは，
ｐ＝７，９３３／４０９，５０２，０７７
＝０．０００１９３７２３０７１１５３０６７６
と算出される。
【００４７】
次に，単位表現抽出手段２２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１８に記録する（ステップＳ２５）。本実施の形態では，「面積は」＋数量表現のパターンをコーパスから取り出し，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・・・ｎ）として，

$${}_nC_r\ p^r\ (1-p)^{n-r}$$

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＝約０．０００１９３７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。

【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。

【００５９】

実際に上記の例でＰの値を計算すると，平方メートル：０．０００００００，ヘクタール：０．０００００００，平方キロメートル：０．０００００００，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：０．０００００００
ヘクタール：０．０００００００
平方キロメートル：０．０００００００
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。

【００６１】

実際，平方メートル：０．０００００００，ヘクタール：０．０００００００，平方キロメートル：０．０００００００，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。

【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】

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：１．０００００００

ヘクタール：１．０００００００

平方キロメートル：１．０００００００

倍：０．９５１４８６７９

分：０．０００８１９８

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：１．０００００００
ヘクタール：１．０００００００
平方キロメートル：１．０００００００
畳：０．９９５８８８６１
平方センチメートル：０．９９９９９９３
が妥当な共起語で，
倍：０．９５１４８６７９
分：０．０００８１９８
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：１．０００００００，ヘクタール：１．０００００００，平方キロメートル：１．０００００００，畳：０．９９５８８８６１，平方センチメートル：０．９９９９９９３は面積の単位表現として利用できるが，倍：０．９５１４８６７９，分：０．０００８１９８，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，ステップＳ５１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。

【００７９】

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。

【００８０】

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。

【００８１】

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４１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。

【００８２】

例えば，「平方メートル」の場合だと，
ｐ＝１７５１０／４０９５０２０７７，ｎ＝７９３３であり，
Ｐ＝１．０００００００００＞０．９９
となって，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。

【００８３】

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Ｓ３の詳細であるステップＳ８１乃至ステップＳ９２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Ｓ８４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Ｓ８６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。

【００８４】

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。

【００８５】

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Ｓ３の詳細であるステップＳ６１乃至ステップＳ７２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Ｓ６４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Ｓ６６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。

【００８６】

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。

【００８７】

超幾何分布とは，
ｈｇ（Ｎ，ｋ，ｎ，ｍ）＝Ｃ（Ｋ，ｍ）×Ｃ（Ｎ－Ｋ，ｎ－ｍ）／Ｃ（Ｎ，ｎ）
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）＝Σｈｇ（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，Σｈｇ（Ｎ，Ｋ，ｎ，ｍ）は，ｍ≧ｋであるｍについてのｈｇ（Ｎ，ｋ，ｎ，ｍ）の合計を意味する。

【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。

【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。

【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。

【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＝ｈｇｓ（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＝ｈｇｓ（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。

【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。

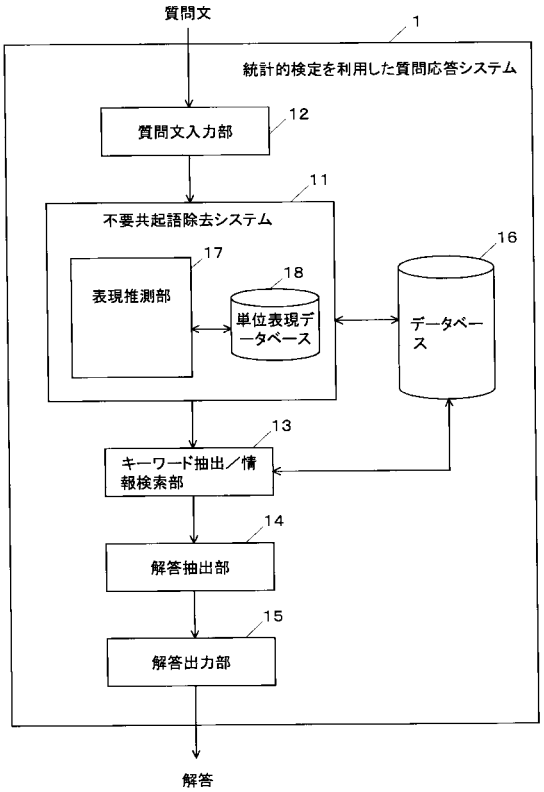
【図１０】不要な単位表現の除去処理フローの詳細の一例を示す図である。

【符号の説明】

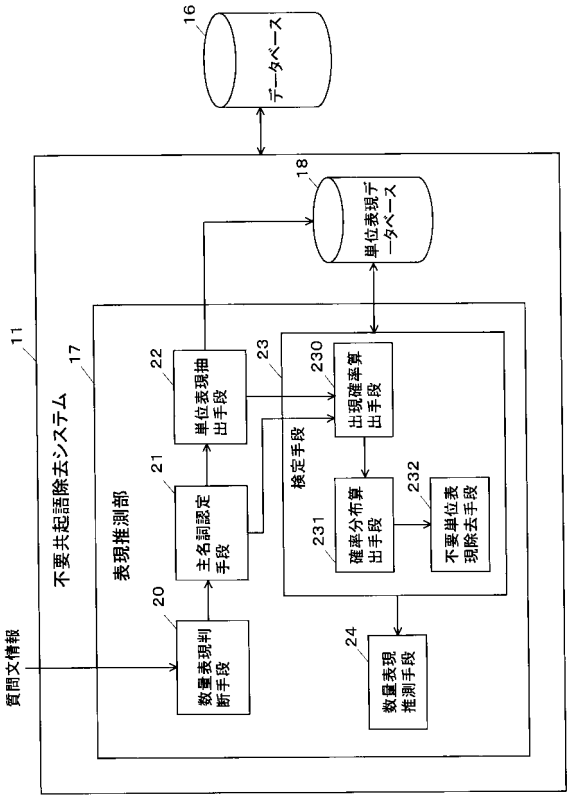
１　統計的検定を利用した質問応答システム
１１　不要共起語除去システム
１２　質問文入力部
１３　キーワード抽出／情報検索部
１４　解答抽出部
１５　解答出力部
１６　データベース
１７　表現推測部
１８　単位表現データベース

２０　数量表現判断手段
２１　主名詞認定手段
２２　単位表現抽出手段
２３　検定手段
２４　数量表現推測手段
２３０　出現確率算出手段
２３１　確率分布算出手段
２３２　不要単位表現除去手段

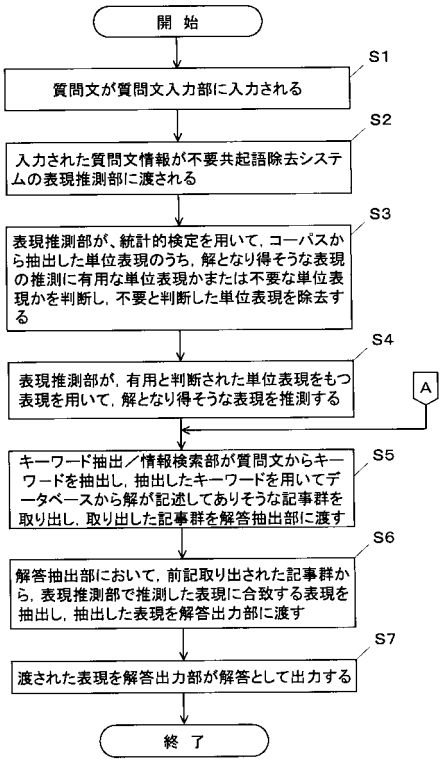
【図１】



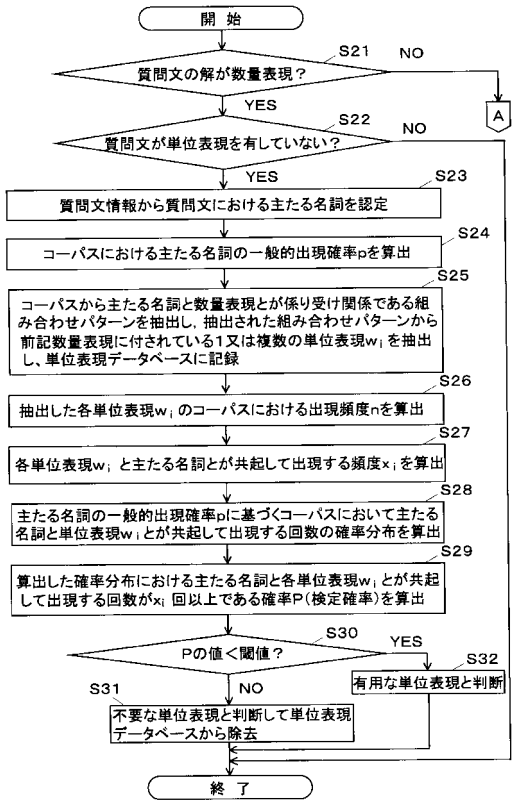
【図２】



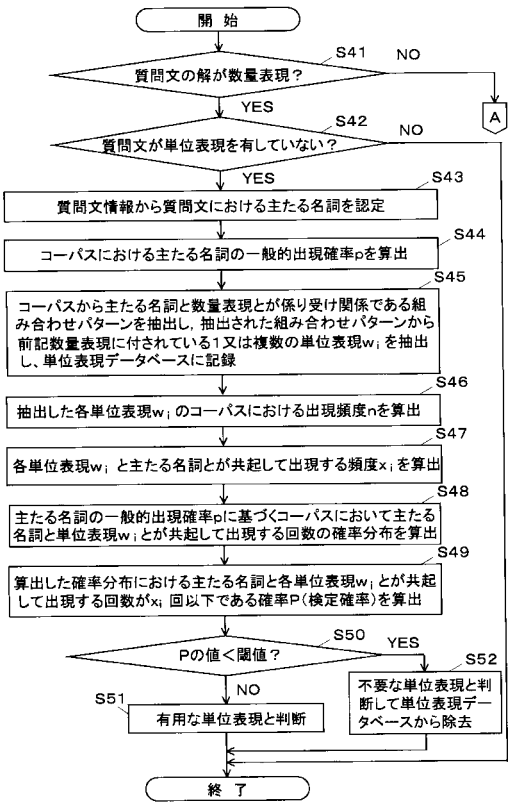
【図３】


開 始
S1
質問文が質問文入力部に入力される
S2
入力された質問文情報が不要共起語除去システムの表現推測部に渡される
S3
表現推測部が、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
S4
表現推測部が、有用と判断された単位表現をもつ表現を用いて、解となり得そうな表現を推測する
A
S5
キーワード抽出／情報検索部が質問文からキーワードを抽出し、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に渡す
S6
解答抽出部において、前記取り出された記事群から、表現推測部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に渡す
S7
渡された表現を解答出力部が解答として出力する
終 了

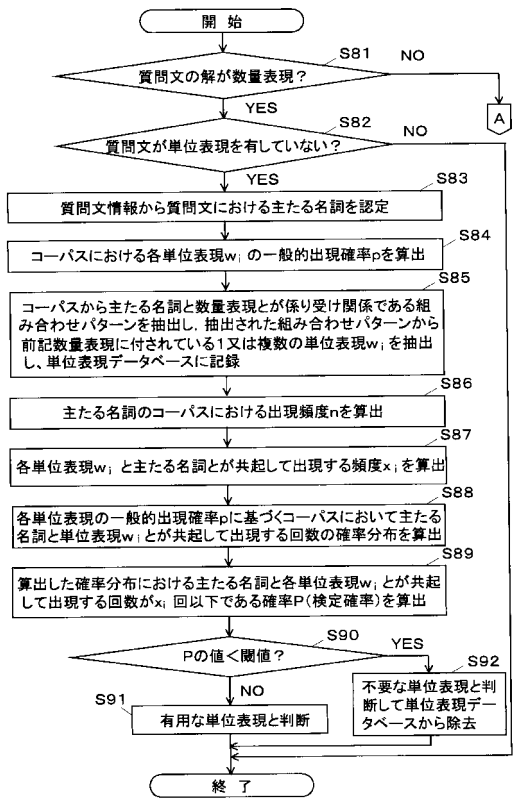
【図４】


開 始
S21
質問文の解が数量表現？
NO
YES
S22
質問文が単位表現を有していない？
NO
A
YES
S23
質問文情報から質問文における主たる名詞を認定
S24
コーパスにおける主たる名詞の一般的出現確率pを算出
S25
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に記録
S26
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
S27
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
S28
主たる名詞の一般的出現確率pに基づくコーパスにおいて主たる名詞と単位表現w_iとが共起して出現する回数の確率分布を算出
S29
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回以上である確率P(検定確率)を算出
S30
Pの値＜閾値？
YES
S32
有用な単位表現と判断
NO
S31
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
終 了

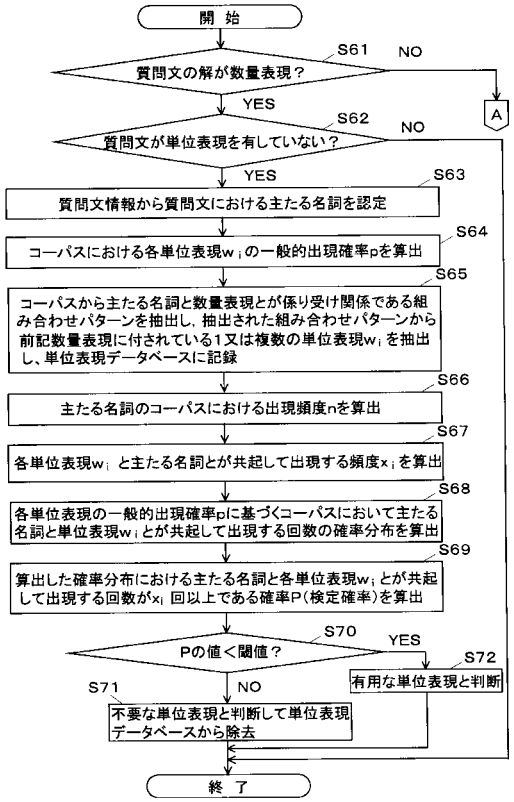
【図５】


開 始
S41
質問文の解が数量表現？
NO
YES
S42
質問文が単位表現を有していない？
NO
A
YES
S43
質問文情報から質問文における主たる名詞を認定
S44
コーパスにおける主たる名詞の一般的出現確率pを算出
S45
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に記録
S46
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
S47
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
S48
主たる名詞の一般的出現確率pに基づくコーパスにおいて主たる名詞と単位表現w_iとが共起して出現する回数の確率分布を算出
S49
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
S50
Pの値＜閾値？
YES
S52
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
NO
S51
有用な単位表現と判断
終 了

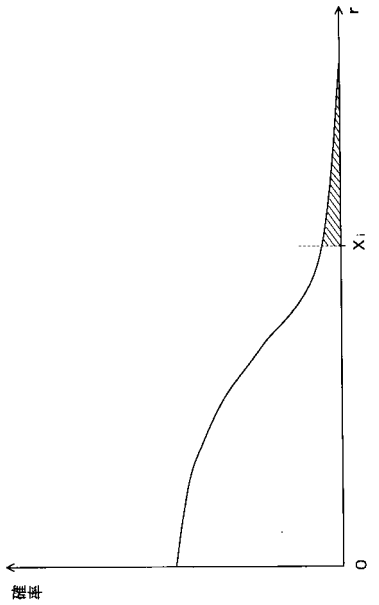
【図６】


開 始
S81
質問文の解が数量表現？
NO
YES
S82
質問文が単位表現を有していない？
NO
A
YES
S83
質問文情報から質問文における主たる名詞を認定
S84
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
S85
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に記録
S86
主たる名詞のコーパスにおける出現頻度nを算出
S87
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
S88
各単位表現の一般的出現確率pに基づくコーパスにおいて主たる名詞と単位表現w_iとが共起して出現する回数の確率分布を算出
S89
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
S90
Pの値＜閾値？
YES
S92
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
NO
S91
有用な単位表現と判断
終 了

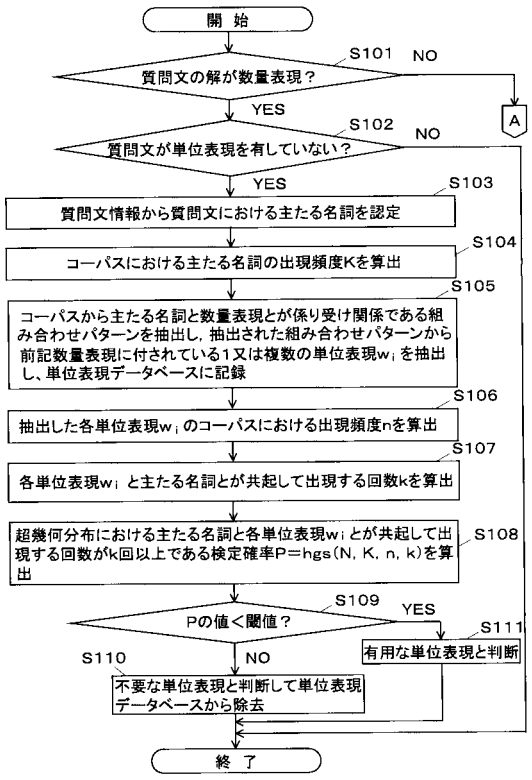
【図７】


開 始
S61
質問文の解が数量表現？
NO
YES
A
S62
質問文が単位表現を有していない？
NO
YES
S63
質問文情報から質問文における主たる名詞を認定
S64
コーパスにおける各単位表現wiの一般的出現確率pを算出
S65
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wiを抽出し、単位表現データベースに記録
S66
主たる名詞のコーパスにおける出現頻度nを算出
S67
各単位表現wi と主たる名詞とが共起して出現する頻度xiを算出
S68
各単位表現の一般的出現確率pに基づくコーパスにおいて主たる名詞と単位表現wiとが共起して出現する回数の確率分布を算出
S69
算出した確率分布における主たる名詞と各単位表現wiとが共起して出現する回数がxi 回以上である確率P(検定確率)を算出
S70
Pの値＜閾値？
YES
S72
有用な単位表現と判断
S71
NO
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
終 了

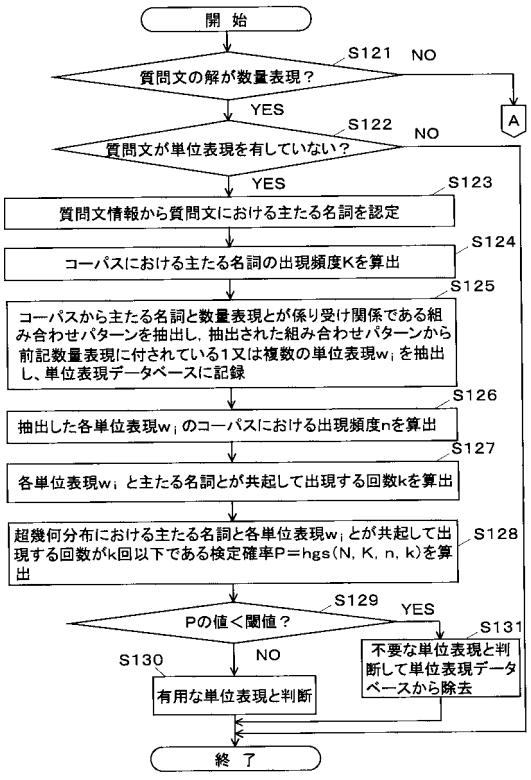
【図８】


P
xi
0
確率

【図９】


開 始
S101
質問文の解が数量表現？
NO
YES
A
S102
質問文が単位表現を有していない？
NO
YES
S103
質問文情報から質問文における主たる名詞を認定
S104
コーパスにおける主たる名詞の出現頻度Kを算出
S105
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wiを抽出し、単位表現データベースに記録
S106
抽出した各単位表現wiのコーパスにおける出現頻度nを算出
S107
各単位表現wi と主たる名詞とが共起して出現する回数kを算出
S108
超幾何分布における主たる名詞と各単位表現wi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＝hgs(N, K, n, k)を算出
S109
Pの値＜閾値？
YES
S111
有用な単位表現と判断
S110
NO
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
終 了

【図１０】


開 始
S121
質問文の解が数量表現？
NO
YES
A
S122
質問文が単位表現を有していない？
NO
YES
S123
質問文情報から質問文における主たる名詞を認定
S124
コーパスにおける主たる名詞の出現頻度Kを算出
S125
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wiを抽出し、単位表現データベースに記録
S126
抽出した各単位表現wiのコーパスにおける出現頻度nを算出
S127
各単位表現wi と主たる名詞とが共起して出現する回数kを算出
S128
超幾何分布における主たる名詞と各単位表現wi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＝hgs(N, K, n, k)を算出
S129
Pの値＜閾値？
YES
S131
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
S130
NO
有用な単位表現と判断
終 了