(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2009 年 9 月 17 日(17.09.2009)

PCT

(10) 国際公開番号
WO 2009/113494 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号: PCT/JP2009/054425

(22) 国際出願日: 2009 年 3 月 9 日(09.03.2009)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2008-060292 2008 年 3 月 10 日(10.03.2008) JP

(71) 出願人 (米国を除く全ての指定国について): 国立大学法人横浜国立大学(NATIONAL UNIVERSITY CORPORATION YOKOHAMA NATIONAL UNIVERSITY) [JP/JP]; 〒2408501 神奈川県横浜市保土ヶ谷区常盤台７９番１号 Kanagawa (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 森 辰則 (MORI, Tatsunori) [JP/JP]; 〒2408501 神奈川県横浜市保土ヶ谷区常盤台７９番１号 国立大学法人横浜国立大学内 Kanagawa (JP). 佐藤 充(SATO, Mitsuru) [JP/JP]; 〒2408501 神奈川県横浜市保土ヶ谷区常盤台７９番１号 国立大学法人横浜国立大学内 Kanagawa (JP). 石下 円香(ISHIOROSHI, Madoka) [JP/JP]; 〒2408501 神奈川県横浜市保土ヶ谷区常盤台７９番１号 国立大学法人横浜国立大学内 Kanagawa (JP).

(74) 代理人: 竹内 三明(TAKEUCHI, Mitsuaki); 〒2510025 神奈川県藤沢市鵠沼石上１丁目４番１３号ロコテラス湘南３０２号室 相州国際特許事務所 Kanagawa (JP).

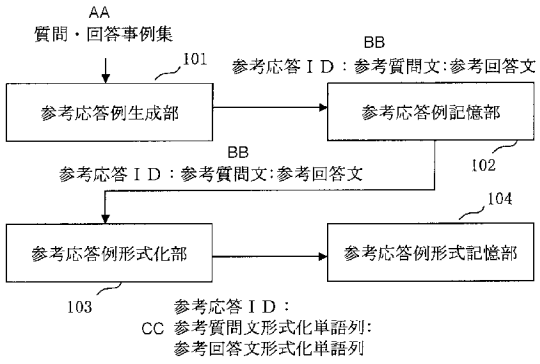
(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY,

[続葉有]

(54) **Title**: QUESTION AND ANSWER SYSTEM WHICH CAN PROVIDE DESCRIPTIVE ANSWER USING WWW AS SOURCE OF INFORMATION

(54) 発明の名称: WWWを情報源として記述的な回答が可能な質問応答システム

[図1]



＜参考応答例準備処理に係る構成＞
DD

AA... QUESTION AND ANSWER SAMPLES
101... REFERENCE ANSWER SAMPLE GENERATING SECTION
BB... REFERENCE ANSWER ID: REFERENCE QUERY SENTENCE: REFERENCE ANSWER SENTENCE
102... REFERENCE ANSWER SAMPLE STORING SECTION
103... REFERENCE ANSWER SAMPLE FORMALIZING SECTION
104... REFERENCE ANSWER SAMPLE FORMAT STORING SECTION
CC... REFERENCE ANSWER ID: STRING OF WORD FOR FORMALIZING REFERENCE QUERY SENTENCE: STRING OF WORD FOR FORMALIZING REFERENCE ANSWER SENTENCE
DD... CONFIGURATION OF PREPARATION PROCESSING OF REFERENCE ANSWER SAMPLE

(57) **Abstract**: Provided is a technology to receive a question expressed by ordinary words, extract candidate answers for the question from among the documents on the WWW, and provide the answers to the user, which reflects formal correlations between questions and answers. A query sentence (1001) is converted into strings of words (1002) for formalizing the query sentence by conducting formalization processing wherein functional words, interrogative words, or predetermined words such as "reasons", "meanings", etc. which tend to be asked are converted into surface expressions and the other content words are converted into word class types. From among the strings of words (1002), key parts (1003) for a query sentence format, each consisting of an interrogative word and a predetermined number of strings of words before and after the interrogative word are extracted. Reference query sentences in question and answer samples are converted in like manner. Then, the aptitude of the format is judged from a degree of similarity of key parts (1004) for a reference query sentence format. Furthermore, reference answer sentences in the question and answer samples are also formalized and a degree of formal correlation with the query sentence is determined.

(57) 要約:

[続葉有]

WO 2009/113494 A1

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

通常の言葉で記された質問を受け付け、その回答候補をWWW上の文書から抽出し、利用者に提示する技術に関し、質問と回答の形式的な相関を反映することを課題とする。質問文１００１を、機能語、疑問詞、あるいは質問の焦点になりやすい「理由」や「意味」などの所定語を表層表現に変換し、その他の内容語を品詞種別に変換する形式化処理により質問文形式化単語列１００２とし、更にその中から疑問詞と所定数の前後の単語列からなる質問文形式要部１００３を抽出する。質問回答事例集の参考質問文も同様に変換し、参考質問文形式要部１００４の類似度により形式の適性を判断する。また、質問回答事例集の参考回答文も形式化し、質問文との形式的な相関度を求める。

明 細 書

WWWを情報源として記述的な回答が可能な質問応答システム

技術分野

[0001] 本発明は、通常の言葉で記された質問を受け付け、その回答候補をWWW上の文書から抽出し、利用者に提示する技術に関する。

背景技術

[0002] WWWを情報源とした質問応答システムが従来研究されている。人名、地名、数量等短い表現が問えるfactoid型、定義や理由等の長い記述が問えるnon－factoid型がある。ここではnon－factoid型に注目する。

[0003] non－factoid型の質問応答における解候補となるテキストの適切性は、「質問文との内容に関する関連性があるか」(観点1)、「質問型に対する回答の仕方(記述スタイル)が適切であるか」(観点2)という二つの観点において計ることができるといわれている。ここで、質問型は質問文が問うている質問の種類(「定義」、「方法」、「理由」など)である。記述スタイルとは、例えば、「理由」を記述するのであれば、「～からである。」「～ために、…」などのように、「理由」を表現するのに適した表現を含む記述様式である。

[0004] 一般的な手法においては、質問型をまず推定してから、それに応じた処理を特に観点2について行う。しかし、質問型の推定の精度の問題や、観点2の判定に利用する手がかり表現を質問型に応じて手作業で準備する必要があるという労力の問題があった。

[0005] これに対して、非特許文献1は、人手による質問応答コミュニティサイトにある大量の質 問・回答事例を学習データとして用いて記述スタイルを獲得することにより、質問の型の推定を行わずに回答を行う学習型のnon－factoid型質問応答の手法を提案している。

[0006] 非特許文献2は、ＦＡＱサイトの質問・回答事例集合から、回答が質問に「書き換え」られる確率を計算し、質問の型に依存しないnon－factoid型質問応答手法を提案している。なお、いずれの手法もＦＡＱなどの質問・回答事例集合は、質問に対する

回答を抽出する対象である情報源ではないことに注意されたい。情報源は別にあって、電子化された新聞記事であったり、WWW上の文書であったりする。

非特許文献1:水野淳太、他1名,「任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討」, 言語処理学会13回年次大会発表論文集(2007), 言語処理学会, 平成19年3月, p. 1002－1005

非特許文献2:ラデゥ ソリカット(Rude Soricut)、他1名, オートマティック クエスチョン アンサーリング ユージング ザ ウェブ(Automatic Question Answering Using the Web), Journalof Information Retrieval- Special Issue on WebInformation Retrieval, November 2006, Vol.9, pp.191-206

発明の開示

発明が解決しようとする課題

[0007] 非特許文献1の方法では、質問の型の推定を陽に行う必要がなく、手がかり表現も質問・回答事例集合から自動的に学習されるという利点がある。しかし、回答選定の柔軟性に問題がある。同手法では、その方法論により、テキストを段落等の予め決められた大きさの単位で切り出したものを観点1により順位付けし、回答候補とした上で、観点2に従い回答になるか否かの判定をするとともに、観点2において順位付けをし直す。ここで、質問応答において、回答の範囲は通常固定ではなく様々であることが普通であることを考えると、この手法では短かったり長かったりと不完全な回答しか得られないことがあると考えられる。さらに、再順位付けにおいては、観点1によらず観点2での並べかえを行うので、観点1である内容の関連性に由来する解候補の重要さは十分に反映されない。

[0008] 非特許文献2の手法でも、回答テキストの大きさを前もって決めておく必要があるとともに、質問の長さに基づいて回答の長さを別途推定する必要がある。さらに、尺度1と尺度2を語の書き換え確率として同時に扱っているために、学習される情報は、内容に纏わるものと記述スタイルに纏わるものが混在している。そのため、学習に利用できる質問・回答事例集合に現れる表現の網羅性が担保されないと精度が低くなると考えられる。

[0009] 本発明においては、質問応答コミュニティサイトにある大量の質問・回答事例集を、

観点2の記述スタイルの適切さを判定するためだけに用い、観点1についての尺度を別途用意して組み合わせることにより、任意の型のnon－factoid型質問応答を行う。また、観点2の記述スタイルの適切さの判定には学習型ではなく、質問・回答事例集の質問側を記述スタイルの類似度に基づいて利用者が与えた質問により検索し、対応する回答事例から動的に回答の記述スタイルに関する情報を取得する。上記のような構成にすることにより、i) 観点1と観点2に関する尺度を独立に設けることができ、なおかつ、それらを同時に考慮できるように統合した一つの評価尺度にすることができる、ii) 使える質問・回答事例集が増えたときには、学習をしなおす必要がなく、単に登録を追加すればよい。

課題を解決するための手段

[0010] 本発明に係る質問応答システムは、

質問文を入力し、検索対象である文書群から質問文の解に適する文を抽出して、応答文として出力する質問応答システムであって、以下の要素を有することを特徴とする

(1)質問文を入力する質問文入力部

(2)文を単語に分割し、単語毎に品詞種別と表層表現を解析し、各単語が機能語である場合、疑問詞である場合、及び質問の焦点になりやすい所定語である場合に、当該単語を質問応答形式に係る単語であると判定し、それ以外の場合に、当該単語を質問応答内容に係る単語であると判定し、質問応答形式に係る単語は品詞種別に変換し、質問応答内容に係る単語は表層表現に変換し、変換した品詞種別あるいは表層表現を単位とした形式化単語列とする形式化処理により、入力された質問文を質問文形式化単語列に変換する質問文形式化部

(3)疑問詞を所定位置に含む所定単語数の形式化単語列を抜き出す質問文形式要部抽出処理により、質問文形式化単語列から質問文形式要部を抽出する質問文形式要部抽出部

(4)参考質問文と参考回答文の対からなる参考応答例を複数記憶する参考応答例記憶部

(5)参考応答例記憶部に含まれる各参考応答例について、参考質問文を前記形式

化処理により参考質問文形式化単語列に変換し、更に参考回答文を前記形式化処理により参考回答文形式化単語列に変換する参考応答例形式化部

(6)変換された参考質問文形式化単語列と参考回答文形式化単語列の対を、参考応答IDに対応付けて複数記憶する応答例形式記憶部

(7)応答例形式記憶部に含まれる各参考質問文形式化単語列について、前記質問文形式要部抽出処理により参考質問文形式要部を抽出し、前記質問文形式要部と比較し、比較結果が同一又は類似の場合に、当該参考質問文形式要部が抽出された参考質問文形式化単語列の参考応答IDを、入力された質問文に形式が相似する相似形式質問文に係る参考応答IDとして特定する相似形式質問文抽出部

(8)特定された相似形式質問文に係る参考応答ID群を、相似形式質問文集合として記憶する相似形式質問文集合記憶部

(9)各相似形式質問文に係る参考応答IDに対応する参考回答文形式化単語列を応答例形式記憶部から取得し、前記形式化単語列の単語数よりも少ない所定単語数の形式化単語列である応答形式要素を、取得した参考回答文形式化単語列から順に抽出し、各応答形式要素について、当該応答形式要素が応答例形式記憶部に含まれる各参考回答文形式化単語列に含まれるか検索し、当該応答形式要素が含まれる参考回答文形式化単語列に係る参考応答ID群を応答形式要素含回答文集合として記憶し、少なくとも相似形式質問文集合と応答形式要素含回答文集合の両方に含まれる参考応答ID群の数、相似形式質問文集合に含まれる参考応答ID群の数、及び応答形式要素含回答文集合に含まれる参考応答ID群の数を用いて、当該応答形式要素を含む参考回答文形式化単語列と相似形式質問文が組み合せられる確率に基づき、当該応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度を算出する応答形式要素相関度算出部

(10)応答形式要素毎に算出された質問形式相関度を記憶する応答形式要素相関度テーブル

(11)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から文書を検索し、検索した文書群に含まれる単語の出現頻度に基づいて、内容として質問文に関連する関連語を抽出するとともに当該関連語の関連度を算

出する関連語生成部

(12)関連語毎に算出された関連度を記憶する関連語テーブル

(13)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から関連文書を検索する検索関連文書検索部

(14)検索された関連文書を、関連文書に含まれる関連文毎に文番号を対応付けて記憶する関連文書記憶部

(15)各関連文について、当該関連文を前記形式化処理により関連文形式化単語列に変換し、関連文形式化単語列から前記応答形式要素を順に抽出し、各応答形式要素の質問形式相関度を応答形式要素相関度テーブルから取得し、更に当該関連文に含まれる各単語の関連語としての関連度を関連語テーブルから取得し、取得した各応答形式要素の質問形式相関度及び各単語の関連度に基づいて、質問文に対する解としての適性を示す文スコアを算出する文スコア算出部

(16)関連文毎の文スコアを文番号に対応付けて記憶する文スコアテーブル

(17)高い適性を示す文スコアの関連文の文番号を解候補として抽出する解候補抽出部

(18)解候補の文番号により特定される関連文を応答文として出力する応答文出力部。

[0011] 更に、前記質問の焦点になりやすい所定語として、少なくとも「理由」、「方法」、「意味」、又は「違い」の何れかを用いることを特徴とする。

[0012] 更に、前記形式化処理は、各単語が参考応答例の中で出現頻度が高い所定の動詞と形容詞である場合にも、当該単語を質問応答形式に係る単語であると判定することを特徴とする。

[0013] 更に、前記質問文形式要部抽出処理により抜き出される形式化単語列は、疑問詞を中心として前後3つ単語を含む合計7つの単語に係る形式化単語列であることを特徴とする。

[0014] 更に、前記相似形式質問文抽出部は、参考質問文形式要部と質問文形式要部に含まれる疑問詞が一致する場合に限り、類似と判定することを特徴とする。

[0015] 更に、前記応答形式要素相関度算出部は、応答形式要素が相似形式質問文に形

式として関連する程度を示す質問形式相関度として、カイ二乗値の平方根を用いることを特徴とする。

[0016] 更に、前記応答形式要素相関度算出部は、応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度として、ダイス係数を用いることを特徴とする。

[0017] 更に、前記応答形式要素相関度算出部は、応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度として、相互情報量を用いることを特徴とする。

[0018] 更に、前記解候補抽出部は、関連文書に含まれる関連文の順に連続する文スコアについて、極大値を示す文スコアの関連文の文番号を解候補とすることを特徴とする。

[0019] 更に、前記解候補抽出部は、前記極大値の所定割合を超える前後の文スコアの関連文の文番号も解候補に含めることを特徴とする。

[0020] 本発明に係るプログラムは、

質問文を入力し、検索対象である文書群から質問文の解に適する文を抽出して、応答文として出力する質問応答システムであって、

参考質問文と参考回答文の対からなる参考応答例を複数記憶する参考応答例記憶部と、

参考質問文形式化単語列と参考回答文形式化単語列の対を、参考応答IDに対応付けて複数記憶するための応答例形式記憶部と、

相似形式質問文に係る参考応答ID群を、相似形式質問文集合として記憶するための相似形式質問文集合記憶部と、

応答形式要素毎に算出された質問形式相関度を記憶するための応答形式要素相関度テーブルと、

関連語毎に算出された関連度を記憶するための関連語テーブルと、

関連文書を、関連文書に含まれる関連文毎に文番号を対応付けて記憶するための関連文書記憶部と、

関連文毎の文スコアを文番号に対応付けて記憶するための文スコアテーブルと、

を有する質問応答システムとなるコンピュータに、以下の手順を実行させることを特徴とする

(1)質問文を入力する質問文入力手順

(2)文を単語に分割し、単語毎に品詞種別と表層表現を解析し、各単語が機能語である場合、疑問詞である場合、及び質問の焦点になりやすい所定語である場合に、当該単語を質問応答形式に係る単語であると判定し、それ以外の場合に、当該単語を質問応答内容に係る単語であると判定し、質問応答形式に係る単語は品詞種別に変換し、質問応答内容に係る単語は表層表現に変換し、変換した品詞種別あるいは表層表現を単位とした形式化単語列とする形式化処理により、入力された質問文を質問文形式化単語列に変換する質問文形式化手順

(3)疑問詞を所定位置に含む所定単語数の形式化単語列を抜き出す質問文形式要部抽出処理により、質問文形式化単語列から質問文形式要部を抽出する質問文形式要部抽出手順

(4)参考応答例記憶部に含まれる各参考応答例について、参考質問文を前記形式化処理により参考質問文形式化単語列に変換し、更に参考回答文を前記形式化処理により参考回答文形式化単語列に変換する参考応答例形式化手順

(5)応答例形式記憶部に含まれる各参考質問文形式化単語列について、前記質問文形式要部抽出処理により参考質問文形式要部を抽出し、前記質問文形式要部と比較し、比較結果が同一又は類似の場合に、当該参考質問文形式要部が抽出された参考質問文形式化単語列の参考応答IDを、入力された質問文に形式が相似する相似形式質問文に係る参考応答IDとして特定する相似形式質問文抽出手順

(6)各相似形式質問文に係る参考応答IDに対応する参考回答文形式化単語列を応答例形式記憶部から取得し、前記形式化単語列の単語数よりも少ない所定単語数の形式化単語列である応答形式要素を、取得した参考回答文形式化単語列から順に抽出し、各応答形式要素について、当該応答形式要素が応答例形式記憶部に含まれる各参考回答文形式化単語列に含まれるか検索し、当該応答形式要素が含まれる参考回答文形式化単語列に係る参考応答ID群を応答形式要素含回答文集合として記憶し、少なくとも相似形式質問文集合と応答形式要素含回答文集合の両

方に含まれる参考応答ID群の数、相似形式質問文集合に含まれる参考応答ID群の数、及び応答形式要素含回答文集合に含まれる参考応答ID群の数を用いて、当該応答形式要素を含む参考回答文形式化単語列と相似形式質問文が組み合せられる確率に基づき、当該応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度を算出する応答形式要素相関度算出手順

(7)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から文書を検索し、検索した文書群に含まれる単語の出現頻度に基づいて、内容として質問文に関連する関連語を抽出するとともに当該関連語の関連度を算出する関連語生成手順

(8)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から関連文書を検索する検索関連文書検索手順

(9)各関連文について、当該関連文を前記形式化処理により関連文形式化単語列に変換し、関連文形式化単語列から前記応答形式要素を順に抽出し、各応答形式要素の質問形式相関度を応答形式要素相関度テーブルから取得し、更に当該関連文に含まれる各単語の関連語としての関連度を関連語テーブルから取得し、取得した各応答形式要素の質問形式相関度及び各単語の関連度に基づいて、質問文に対する解としての適性を示す文スコアを算出する文スコア算出手順

(10)高い適性を示す文スコアの関連文の文番号を解候補として抽出する解候補抽出手順

(11)解候補の文番号により特定される関連文を応答文として出力する応答文出力手順。

[0021] 本発明に係る質問応答システムは、

参考質問文と参考回答文の対を参考文とし、該参考文のうち少なくとも参考質問文に対して記述スタイルを一般化する形式化処理を行う参考文形式化部と、

前記参考文形式化部において形式化された形式化参考文を記憶する参考文記憶部と、

入力質問文の記述スタイルを一般化する形式化処理を行う入力質問文形式化部と、

前記入力質問文形式化部において形式化された形式化入力質問文と類似する形式を有する前記形式化参考文を探索し、該形式化参考文に含まれる参考回答文を前記参考文記憶部から抽出する参考回答文抽出部と、

前記参考回答文と、前記入力質問文をWebサーチエンジンで検索した結果得られたWeb文書である検索Web文書との間の記述スタイルの適合性を評価する記述スタイル評価部と、

前記検索Web文書と、前記入力質問文との間の内容の関連性を評価する関連性評価部と、

前記記述スタイル評価部により前記参考回答文と記述スタイルの適合性があると評価され、かつ、前記関連性評価部により前記入力質問文の内容と関連があると評価された検索Web文書に対してスコア付け処理を行うスコア処理部と、

該スコアに基づいて、前記入力質問文に対する回答文を出力する回答文出力部を有することを特徴とする。

[0022] 本発明に係るプログラムは、質問応答システムとなるコンピュータに、

参考質問文と参考回答文の対を参考文とし、該参考文のうち少なくとも参考質問文に対して記述スタイルを一般化する形式化処理を行う参考文形式化手順と、

前記参考文形式化手順において形式化された形式化参考文を記憶する参考文記憶手順と、

入力質問文の記述スタイルを一般化する形式化処理を行う入力質問文形式化手順と、

前記入力質問文形式化手順において形式化された形式化入力質問文と類似する形式を有する前記形式化参考文を探索し、該形式化参考文に含まれる参考回答文を抽出する参考回答文抽出手順と、

前記参考回答文と、前記入力質問文をWebサーチエンジンで検索した結果得られたW

eb文書である検索Web文書との間の記述スタイルの適合性を評価する記述スタイル評価手順と、

前記検索Web文書と、前記入力質問文との間の内容の関連性を評価する関連性

評価手順と、

前記記述スタイル評価手順により前記参考回答文と記述スタイルの適合性があると評価され、かつ、前記関連性評価手順により前記入力質問文の内容と関連があると評価された検索Web文書に対してスコア付け処理を行うスコア処理手順と、

該スコアに基づいて、前記入力質問文に対する回答文を出力する回答文出力手順を実行させることを特徴とする。

発明の効果

[0023] 本発明によれば、本発明においては、質問応答コミュニティサイトにある大量の質問・回答事例集を、観点2の記述スタイルの適切さを判定するためだけに用い、観点1についての尺度を別途用意して組み合わせることにより、任意の型のnon－factoid型質問応答を行う。また、観点2の記述スタイルの適切さの判定には学習型ではなく、質問・回答事例集の質問側を記述スタイルの類似度に基づいて利用者が与えた質問により検索し、対応する回答事例から動的に回答の記述スタイルに関する情報を取得するので、観点1と観点2に関する尺度を独立に設けることができ、なおかつ、それらを同時に考慮できるように統合した一つの評価尺度にすることができる。使える質問・回答事例集が増えたときには、学習をしなおす必要がなく、単に登録を追加すればよい。

発明を実施するための最良の形態

[0024] 実施の形態1.

まず、参考応答例を準備する動作について説明する。参考応答例は、質問応答システムによる質問応答のための学習用データであって、例えば質問とそれに対応する回答の事例を集めた既存の質問・回答事例集合を用いる。

[0025] 図1は、参考応答例準備処理に係る構成を示す図である。質問応答システムは、質問・回答事例集合から学習用データとしての参考応答例を生成する参考応答例生成部101、生成した参考応答例（参考質問文と参考回答文の対）に参考応答IDを対応付けて記憶する参考応答例記憶部102、参考応答例に含まれる参考質問文と参考回答文を所定の手順に従って形式化する参考応答例形式化部103、形式化された参考質問文形式化単語列と参考回答文形式化単語列の対を参考応答IDに対応

付けて記憶する参考応答例形式記憶部104を有している。

[0026] 図2は、参考応答例準備処理フローを示す図である。この例では、Webコミュニティサービスの利用者同士でなされた質問・回答事例集合を用いる。従って、参考応答例生成部101による参考応答例生成部(S201)では、一つの質問に複数の回答が対応する場合に、質問者が最良回答として選んだ回答を参考回答文とする。また、質問が1文であって回答文にURLを含まない質問・回答のみを選択して、参考応答例として参考応答IDに対応付けて参考質問文及び参考回答文として参考応答例記憶部102に記憶させる。

[0027] 図3は、参考応答例記憶部の構成例を示す図である。参考応答毎にレコードを設け、参考応答ID351と、参考質問文352と、参考回答文353との項目を対応付けて記憶するように構成されている。

[0028] 参考応答例形式化部103による参考応答例形式化処理(S202)では、質問及び回答の内容的な意義を排除して質問及び回答としての形式的な意義のみを有する情報(形式化単語列と呼ぶ。)に変換する。

[0029] 図4は、参考応答例形式化処理フローを示す図である。参考応答例記憶部102に含まれる参考応答例毎に以下の処理を繰り返す(S401)。参考質問文を質問文形式化処理(図6)し、参考質問文形式化単語列を得て(S402)、更に参考回答文を回答文形式化処理(図6)し、参考回答文形式化単語列を得る(S403)。そして、これらを参考応答IDに対応付けて参考応答例形式記憶部104に記憶させる。これをすべての参考応答例について処理する(S404)。

[0030] 図5は、参考応答例形式記憶部の構成例を示す図である。参考応答毎にレコードを設け、参考応答ID551と、参考質問文形式化単語列552と、参考回答文形式化単語列553との項目を対応付けて記憶するように構成されている。このように、形式化単語列は、形式に係る語の表層表現(読み)と、内容に係る語の品詞種別を単語の列として並べた構成となっている。

[0031] ここで、具体的な質問文形式化処理および回答文形式化処理について説明する。図6は、質問文形式化処理／回答文形式化処理フローを示す図である。この例では、質問文形式化処理と回答文形式化処理は共通である。まず、対象文(参考質問文

又は参考回答文、あるいは後述する質問文又は関連文)を形態素解析し、単語毎の品詞種別と表層表現を得る(S601)。そして、単語毎に(S602)、質問応答特性判定処理(S603、図7)により、当該単語の特性を判定する。質問応答形式に係る単語であると判定された場合には、当該単語の品詞種別を、対象文形式化単語列(参考質問文形式化単語列又は参考回答文形式化単語列、あるいは後述する質問文形式化単語列又は関連文形式化単語列)に追加する(S604)。一方、質問応答内容に係る単語と判定された場合には、当該単語の表層表現を、同様に対象文形式化単語列に追加する(S605)。尚、単語間には区切りの記号を入れて単語の単位を識別できるようにする。また、句読点などの記号も単語として扱う。これらの処理をすべての単語に対して行う(S606)。

[0032] ここで、前述の質問応答特性判定処理(S603)について説明する。図7は、質問応答特性判定処理フローを示す図である。当該単語の品詞が助詞、助動詞等の機能語であるか判断し(S701)、機能語である場合には質問応答形式に係る単語と判定する(S706)。また、当該単語が疑問詞か判定する(S702)。例えば、疑問詞となる代名詞としては、ナニ、ドコ、ダレ、ナン、ドチラ、ドレ、ドッチ、イツ、ドナタ、イクツ、ドッカ、イズレ、ナアニ等がある。疑問詞となる連体詞としては、ドノ、ドンナ、ドウイウ、イカナル等がある。また、疑問詞となる副詞としては、ドウ、ナゼ、ドウシテ、イクラ、イツノマニ等がある。その他にも、ッテナ、ナニモノ等ある。これらの品詞と表層表現の組み合せによって判定する。そして、疑問詞である場合には質問応答形式に係る単語と判定する(S706)。また、内容語であっても、所定の質問の焦点となりやすい単語であるかを判定し(S703)、その所定の単語である場合には、質問応答形式に係る単語と判定する(S706)。例えば、「理由」、「方法」、「意味」、「違い」等が質問の焦点となりやすい単語である。また、参考応答例の中で出現頻度が高い動詞と形容詞も予め特定しておき、その所定の頻出する単語であるかも判定し(S704)、所定の頻出する単語の場合には、質問応答形式に係る単語と判定する(S706)。そして、その他の内容語は、質問応答内容に係る単語と判定する(S706)。S704のステップは省略することもできる。

[0033] 上述の動作により、質問応答システムにより参考応答例の準備を事前に行っておく

。次に、実際の質問応答の動作について説明する。

[0034] 図8は、質問応答処理フローを示す図である。図に示すように順次、質問文を入力する質問文入力処理(S801)と、質問文を形式化する質問文形式化処理(S802)と、形式化した質問文から質問形式としての要部を抽出する質問文形式要部抽出処理(S803)と、参考応答例から質問形式の要部が相似(同一あるいは類似)する参考質問文を抽出する相似形式質問文抽出処理(S804)と、相似する参考質問文に対する参考応答文を形式化し、その形式化された参考回答文に含まれる単語列からなる応答形式要素について、当該相似する参考質問文との相関度を算出する応答形式要素相関度算出処理(S805)と、質問文に内容的に関連する関連語を生成する関連語生成処理(S806)と、質問文に内容的に関連する関連文書を検索する関連文書検索処理(S807)と、関連文書に含まれる文(関連文と呼ぶ)毎に、質問文に対する応答としての適性の程度を文スコアとして算出する文スコア算出処理(S808)と、文スコアに基づいて応答解となる候補範囲を抽出する解候補抽出処理(S809)と、抽出した解候補を応答文として出力する応答文出力処理(S810)を行う。

[0035] 以下ではこれらの動作を、質問文入力処理(S801)から応答形式要素相関度算出処理(S805)の前半動作と、関連語生成処理(S806)から応答文出力処理(S810)の後半動作に分けて説明する。

[0036] 図9は、質問文入力から応答形式要素相関度計算までの処理に係る構成を示す図である。質問応答システムは、前述の参考応答例形式記憶部104の他、質問文入力処理(S801)を行う質問文入力部901、入力された質問文を記憶する質問文記憶部902、質問文形式化処理(S802)を行う質問文形式化部903、形式化された質問文形式化単語列を記憶する質問文形式記憶部904、質問文形式要部抽出処理(S803)を行う質問文形式要部抽出部905、抽出された質問文形式要部を記憶する質問文形式要部記憶部906、相似形式質問文抽出処理(S804)を行う相似形式質問文抽出部907、抽出された相似形式の参考質問文の参考応答IDを集合として記憶する相似形式質問文集合記憶部908、応答形式要素相関度算出処理(S805)を行う応答形式要素相関度算出部909、応答形式要素を含む参考回答文の参考応答IDを集合として記憶する応答形式要素含回答文集合記憶部910、応答形式要素が相

似形式質問文に形式的に関連する程度である質問形式相関度を記憶する応答形式要素相関度テーブル911を有している。

[0037] ここで、参考応答例記憶部102の参考質問文の中から、質問文と形式が相似する参考質問文を抽出する手順の概要を説明する。図10は、質問文と参考質問文の比較例を示す図である。前述の通り、各参考質問文1006は予め質問文形式化処理により参考質問文形式化単語列1005に変換しておく。そして、質問文1001が入力されると、これを同様に質問文形式化処理し、質問文形式化単語列1002を得る。更に、質問文形式化単語列1002から疑問詞(この例では、ナニ)を中心とする質問文の形式としての要部である質問文形式要部1003を抽出する。そして、各参考質問文形式化単語列1005について、同様に参考質問文形式要部1004を抽出して、それぞれを比較する。比較結果が完全一致する場合に、質問文形式が同一であると判断する。また、疑問詞が一致する部分一致の場合には、一致の程度に従って類似すると判断する。

[0038] 質問文入力処理(S801)では、操作者の操作等により入力された質問文を質問文記憶部902に記憶させ、質問文形式化処理(S802)では、前述の質問文形式化処理(図7)により質問文記憶部902に記憶している質問文を形式化して、質問文形式化単語列を生成し、質問文形式記憶部904に記憶させる。

[0039] 次に、質問文形式要部抽出処理(S803)について詳述する。図11は、質問文形式要部抽出処理フローを示す図である。まず、質問文形式化単語列中の疑問詞を特定する(S1101)。疑問詞は、前述の通り予め定められた品詞種別と表層表現により特定することができる。そして、疑問詞の前の3単語と後の3単語を含む7単語から成る単語列を抽出し、質問文形式要部とする(S1102)。この例では、所定の疑問詞前単語数を3とし、所定の疑問詞後単語数を3として説明する。

[0040] 次に、相似形式質問文抽出処理(S804)について詳述する。図12は、相似形式質問文抽出処理フローを示す図である。参考応答例形式記憶部104に含まれる参考応答例毎に以下の処理を繰り返す(S1201)。まず、当該参考応答例の参考質問文形式化単語列から参考質問文形式要部を抽出する(S1202)。抽出の手順は、前述の質問文形式要部抽出処理(S803:図11)と同様である。そして、抽出した参考

質問文形式要部を質問文形式要部記憶部906に記憶している質問文形式要部と比較する。中央の疑問詞が一致しない場合には(S1203)、相似度を0とし(S1205)、相似しないものとして扱う。中央の疑問詞が一致する場合には(S1203)、各位置の単語の一致数をカウントし、相似度とする(S1204)。つまり、位置と単語の両方が一致した場合を1として、中央以外の6つの位置の一致した数を合計する。このとき、単語が表層表現のときには表層表現が一致する場合、単語が品詞種別のときには品詞種別が一致する場合に、単語が一致するとして処理する。そして、すべての参考応答例について処理を終えると(S1206)、計数した相似度の高い順に従って、相似形式質問文を選択する(S1207)。選択数(1又は2以上)を予め設定しておき、所定の選択数に達するまで選択する方法や、選択基準となる相似度の下限(単語全数である7又は6以下)を予め設定しておき、所定の相似度下限以上の相似形式質問文を選択する方法が考えられる。そして、選択した相似形式質問文を識別するための参考応答IDを相似形式質問文集合記憶部908に記憶させる。これにより、相似形式質問文を要素とする相似形式質問文集合について、参考応答IDを識別子として取り扱うことができる。

[0041] 次に、応答形式要素相関度算出処理(S805)について詳述する。図13は、応答形式要素相関度算出処理フローを示す図である。相似形式質問文集合記憶部908で記憶している相似形式質問文の参考応答IDを読み出し、参考応答例形式記憶部104からその参考応答IDに対応する参考回答文形式化単語列を読み出す。その参考回答文形式化単語列から、順に連続する2単語を抽出し、応答形式要素とする。この処理を相似形式質問文集合記憶部908で記憶している各相似形式質問文に対して行う(S1301)。尚、2単語が重複する場合には、省略して構わない。応答形式要素とは、回答文形式単語列よりも小さい単位の形式化された単語列である。この例では、単語数を2とする。そして、抽出した応答形式要素毎に以下の処理を繰り返す(S1302)。

[0042] まず、参考応答例形式記憶部104から当該応答形式要素を含む参考回答文形式化単語列を検索し、その参考応答IDを応答形式要素含回答文集合として応答形式要素含回答文集合記憶部910に記憶させる(S1303)。

[0043] 次に、当該応答形式要素と相似形式質問文の相関度をカイ二乗検定により求める。この場合のカイ二乗値を算出する式を示す。

[0044] [数1]

$$\chi^2(b,A)=\frac{n\cdot\left(|A\cap B|\cdot|\bar{A}\cap\bar{B}|-|\bar{A}\cap B|\cdot|A\cap\bar{B}|\right)^2}{|A|\cdot|\bar{A}|\cdot|B|\cdot|\bar{B}|}$$

[0045] nは、全参考応答例数であり、Aは、相似形式質問文からなる参考応答例の集合であり、Bは、応答形式要素含回答文からなる参考応答例の集合である。相似形式質問文と応答形式要素含回答文が共起する頻度に基づいて、両者の相関を求めることができる。

[0046] 処理としては、まず式中の各所定集合の要素数を算出する(S1304)。全参考応答例数nとして、参考応答例形式記憶部104に参考応答ID数を計数する。また分母の各項について、集合Aの要素数として相似形式質問文集合記憶部908に含まれる参考応答ID数を計数し分母第1項値とし、Aの余集合の要素数として全参考応答例数nから順に集合Aの要素数を減じて差を求め分母第2項値とし、集合Bの要素数として応答形式要素含回答文集合記憶部910に含まれる参考応答ID数を計数し分母第3項値とし、Bの余集合の要素数として全参考応答例数nから集合Bの要素数を減じて差を求め分母第4項値とする。更に分子括弧内の各項について、集合Aと集合Bの積集合の要素数として相似形式質問文集合記憶部908と参考応答例形式記憶部104に共に含まれる参考応答ID数を計数し分子括弧内第1項値とし、Aの余集合とBの余集合の積集合の要素数として相似形式質問文集合記憶部908と参考応答例形式記憶部104のいずれにも含まれない参考応答ID数を計数し分子括弧内第2項値とし、Aの余集合と集合Bの積集合の要素数として相似形式質問文集合記憶部908に含まれず参考応答例形式記憶部104に含まれる参考応答ID数を計数し分子括弧内第3項値とし、集合AとBの余集合の積集合の要素数として相似形式質問文集合記憶部908に含まれ参考応答例形式記憶部104に含まれない参考応答ID数を計数し分子括弧内第4項値とする。

[0047] 次に、各集合の要素数と全参考応答例数からカイ二乗値を算出する(S1305)。まず、分母第1項値と分母第2項値と分母第3項値と分母第4項値を積算し、分母値を求める。次に、分子括弧内第1項値と分子括弧内第2項値を積算し分子括弧内前項値を求め、分子括弧内第3項値と分子括弧内第4項値を積算し分子括弧内後項値を求め、分子括弧内前項値と分子括弧内後項値の差を求め、差の二乗値に全参考応答例数nを乗じて分子値を求める。最語に、分子値を分母値で割って、カイ二乗値とする。また、カイ二乗値の二乗根を算出し、当該二乗根を質問形式相関度として当該応答形式要素に対応付けて記憶する(S1306)。この処理をすべての応答形式要素について行う(S1307)。

[0048] 上述の処理により応答形式要素相関度テーブル911が生成される。図14は、応答形式要素相関度テーブルを示す図である。応答形式要素毎にレコードを設け、応答形式要素1451と、カイ二乗値1452と、質問形式相関度1453との項目を対応付けて記憶するように構成されている。この例では、カイ二乗値も記憶させているが省略することもできる。

[0049] 図14の例は、図10の1003に示したタ__リユウ__ハ__ナニ__デス__カ__<記号,句点,*,*>の質問文形式要部に対して、タ__リユウ、タ__カラ、リユウ__ハなどの応答形式要素1451が、形式的に相関が高いということを示している。

[0050] 続いて、図8に示した関連語生成処理(S806)から応答文出力処理(S810)の後半動作について説明する。

[0051] 図15は、関連語生成から応答文出力までの処理に係る構成を示す図である。質問応答システムは、前述の質問文記憶部902と応答形式要素相関度テーブル911の他、関連語生成処理(S806)を行う関連語生成部1501、質問文に内容的に関連する関連語を内容的な関連度と共に記憶する関連語テーブル1502、関連文書検索処理(S807)を行う関連文書検索部1503、質問文に内容的に関連する関連文書を記憶する関連文書記憶部1504、文スコア算出処理(S808)を行う文スコア算出部1505、関連文書に含まれる関連文毎に、質問文に対する応答としての適性の程度を文スコアとして記憶する文スコアテーブル1506、解候補抽出処理(S809)を行う解候補抽出部1507、文スコアに基づいて応答解と判定された候補範囲を記憶する解

候補記憶部1508、応答文出力処理(S810)を行う応答文出力部1509を有している。

[0052] 次に、関連語生成処理(S806)について詳述する。図16は、関連語生成処理フローを示す図である。まず、質問文記憶部902に記憶している質問文から複数のキーワードを抽出する。この例では、質問文から複合語を含む動詞・形容詞のキーワードを抽出する(S1601)。そして、順次キーワードを組み合せてクエリを生成する。この例では、3つのキーワード組合せのAND条件からなるクエリを生成する。そしてクエリ毎に以下の処理を繰り返す(S1602)。当該クエリを入力してWeb検索し、検索結果の要約であるスニップ集合を得る(S1603)。尚、質問応答システムはインターネットに接続しており、Web検索サイトを介してWeb上のサイト、HTML文書、及びその他のコンテンツを検索できるように構成されている。そしてスニップ集合に含まれる単語(内容語に限る。以下、関連語候補と呼ぶ。)を抽出し、関連語候補の内容的な関連度を算出し、関連語を特定する。この例における内容的な関連度の算出式を以下に示す。

[0053] [数2]

$$T(w_j) = \max_i \frac{freq(w_j, i)}{n_i}$$

[0054] 式中、w_jは、各単語(各関連語候補)であり、q_iは、各クエリであり、n_iは、各クエリq_iに対して得られたスニップの件数であり、freq(w_j, i)は、各単語w_jの各クエリq_iに対して得られたスニップ集合中でのスニップ頻度であり、T(w_j)は、各単語w_jの内容的な関連度である。

[0055] 関連語候補毎に以下の処理を繰り返す(S1604)。当該関連語候補のスニップ集合中における頻度を算出し、当該スニップ頻度をスニップ数で割り、正規化スニップ頻度とする(S1605)。すべての関連語候補について正規化スニップ頻度を求める(S1606)。この処理を、予定しているすべてのキーワード組合せについて行った時点で(S1607)、関連語候補毎に正規化スニップ頻度同士を比較し、最大の正規化ス

ニップ頻度を内容的な関連度とする。内容的な関連度が所定閾値以上の場合に、当該関連語候補を関連語と判定し、関連語とその関連度を関連語テーブル1502に記憶させる(S1608)。尚、閾値による判定を行なわずに、すべての関連語候補を関連語とした扱う形態も有効である。

[0056] 図17は、関連語テーブルの構成例を示す図である。関連語毎にレコードを設け、関連語1751と関連度1752との項目を対応付けて記憶するように構成されている。

[0057] 図17の例は、図10の1001に示した、「琉球王国のグスク及び関連遺産群」が正解遺産に登録された理由は何ですか、という質問文に対して、2000、沖縄、文化などの関連語1751が、内容的に関連が高いということを示している。

[0058] 次に、関連文書検索処理(S807)について詳述する。図18は、関連文書検索処理フローを示す図である。まず、質問文記憶部902に記憶している質問文から複数のキーワードを抽出する。この例では、質問文から複合語を含む動詞・形容詞のキーワードを抽出する(S1801)。そして、順次キーワードを組み合せてクエリを生成する。この例では、所定数のキーワード組合せのAND条件からなるクエリを生成する。そしてクエリ毎に以下の処理を繰り返す(S1802)。当該クエリを入力してWeb検索し、検索結果の各文書のURLを得る(S1803)。そして当該URLからHTML文書をダウンロードし(S1804)、HTML文書をプレーンテキストに変換し、関連文書とする(S1805)。所定のキーワード組合せについて処理すると(S1806)、プレーンテキストに変換した各関連文書に含まれる関連文毎に、関連文書記憶部1504の全体を通した文番号を割り当てて順次記憶する(S1807)。

[0059] 次に、文スコア算出処理(S808)について説明する。この処理では、関連文毎に、質問文に対する内容的な関連度と形式的な相関度を考慮した応答適性を判定する。

[0060] 図19は、文スコア算出処理フローを示す図である。関連文書記憶部1504で記憶している各関連文書に含まれる関連文を順に特定し、当該関連文毎に以下の処理を繰り返す(S1901)。内容評価項算出処理(S1902)では、内容的な関連度に基づく評価を行ない、形式評価項算出処理(S1903)では、形式的な関連度に基づく評価を行なう。ここで、総合的な評価指標となる文スコアの算出式を示す。

[0061] [数3]

$$\mathrm{Score}(S_i) = \frac{\left\{\sum_{j=1}^{n} T(w_{ij})\right\}^{\alpha} \cdot \left\{\sum_{k=1}^{m} \sqrt{\chi^2(b_{ik})}\right\}^{1-\alpha}}{\log(1+|S_i|)}$$

[0062] 式中、S_iは、各関連文であり、w_{ij}は、各関連文S_iに含まれる各単語であり、nは、関連文S_i中の単語w_{ij}の異なり数であり、b_{ik}は、各関連文Siに含まれる各応答形式要素(この例では、形式化された2単語)であり、mは、関連文S_i中の応答形式要素b_{ik}の異なり数である。Tは、前述と同様に関連語の関連度であり、カイ二乗値の平方根は、応答形式要素の質問形式相関度であり、Score(S_i)は、各関連文S_iの文スコアである。

[0063] 正確な適性を得るためには、関連文内の単語や応答形式要素に関する密度を考慮して文の長さで評価値を割る必要がある。文の長さとして単純に単語数を用いることもできるが、この例では特に、短い文は回答として不適切である場合が多いことを考慮して、短い文の適性を下げる意味で文の長さとして単語数の対数を用いている。

[0064] その為、当該関連文の長さ(1+関連文中単語数)の対数を算出し(S1904)、内容要評価項と形式評価項の積算し、当該積を関連文の長さの対数で除算し、文スコアを求める(S1905)。

[0065] ここで、前述の内容評価項算出処理(S1902)と形式評価項算出処理(S1903)について詳述する。

[0066] 図20は、内容評価項算出処理フローを示す図である。まず、当該関連文中に含まれる単語(内容語に限る)を特定し(S2001)、各単語の関連語としての関連度を関連語テーブル1502から取得する。そして、全ての単語の関連度を加算し、関連度総和を求める(S2002)。更に、関連度総和のα乗を算出して内容評価項値とする(S2003)。この累乗計算に用いる定数αは、0から1の値であって、評価全体に対する内容評価の重みを示している。大きい値ほど内容に対する重みが増す。例えば0の場合は、形式評価のみの文スコアを得ることになり、1の場合には内容評価のみの文スコアを得ることになる。尚、この評価重みのα値は、質問応答システムとして予め設定

し、事前に記憶されている値を用いる方式や、質問文に対する応答文の生成の際に、操作者が設定する方式が考えられる。いずれの場合にも、評価重み値を入力する評価重み入力部と、評価重み値を記憶する評価重み記憶部を有し、本処理はその評価重み値を読み出してα値として累乗算出に用いる。

[0067] 図21は、形式評価項算出処理フローを示す図である。当該関連文を前述と同様に回答文形式化処理(図6)し、関連文形式化単語列を得る(S2101)。そして、関連文形式化単語列から、順に連続する2単語を抽出し、応答形式要素とする(S2102)。各応答形式要素の応答形式相関度を応答形式要素相関度テーブル911から取得し、全ての応答形式要素の質問形式相関度を加算し、質問形式相関度総和を求める(S2103)。そして、質問形式相関度総和の1－α(形式重み)乗を算出して、形式評価項値とする(S2104)。

[0068] 図19に示すように、当該文番号に対応付けて、求めた文スコアを文スコアテーブル1506に記憶し(S1906)、すべての関連文について処理して終了する(S1907)。

[0069] これにより、関連文毎の文スコアが得られる。文スコアテーブル1506で記憶する。図22は、文スコアの分布例を示す図である。図に示すように、文スコアが高い領域が存在する。次に、これを解候補2201, 2202, 2203として抽出する解候補抽出処理(S809)を行なう。

[0070] 図23は、解候補抽出処理フローを示す図である。文スコアテーブル1506に含まれる文番号に従って、関連文毎に以下の処理を繰り返す(S2301)。文スコアが極大値か判定する(S2302)。このとき、前後の文の文スコアよりも大きい場合に極大値とする。極大値である場合には、当該関連文の文番号を文スコア極大値に対応付けて、解候補記憶部1508の解候補の先頭文番号と末尾文番号に記憶する(S2303)。解候補記憶部1508は、解候補毎に、解候補範囲となる先頭文番号と末尾文番号、及び文スコア極大値を対応付けて記憶するように構成されている。更に、順次、前の関連文の文スコアが極大値の1／2以上であるか判定し、1／2以上である場合には解候補の先頭文番号を当該前の関連文の文番号に改める(S2304)。1／2より小さい場合には、その時点で本ステップを終了する。また、順次、後の関連文の文スコアが極大値の1／2以上であるか判定し、1／2以上である場合には解候補の末尾文番号

を当該後の関連文の文番号に改める(S2305)。1／2より小さい場合には、その時点で本ステップを終了する。上述の処理をすべての文について行なう。(S2306)。1／2は、所定の割合の例である。

[0071] 最後に、応答文出力処理(S810)について詳述する。図24は、応答文出力処理フローを示す図である。文スコア極大値の大きい順に、解候補を特定し(S2401)、解候補の先頭文番号から末尾文番号までの関連文書記憶部1504から関連文を得る(S2402)。そして、関連文群を応答文として出力する(S2403)。出力する応答文数が定められている場合には、応答文数分の処理を繰り返す(S2404)。また、出力する応答文量が定められている場合には、応答文量に至るまで処理を繰り返す。あるいは、操作者の指示により応答文を切りかえる場合には、指示に従って上述の処理を行なう。

[0072] 図25は、正解と不正解の例を示す図である。正解2501及び正解2502は、図10の質問文1001に対して得られた応答文の例である。不正解2503は、形式に関する評価が低いために応答文とならなかった例である。

[0073] 図25に示したように、形式的な適性を考慮することにより、内容的な適性も向上することがわかる。

[0074] 実施の形態2.

上述の形態では、質問文形式要部として疑問詞の前後3単語を含む7単語を抽出したが、他の単語数することもできる。

[0075] 前の3単語は、所定の疑問詞前単語数を3とした例であり、他に、1、2、4、5、6以上とすることもできる。また、後の3単語は、所定の疑問詞後単語数を3とした例であり、他に、1、2、4、5、6以上とすることもできる。また、所定前単語数と所定後単語数は同数に限らず、異なる数であっても構わない。

[0076] 実施の形態3.

上述の形態では、応答形式要素数として2単語抽出したが、他の単語数することもできる。連続する3単語、4単語、5単語以上とすることもできる。

[0077] 応答形式要素数は、質問文形式要部の単語数よりも小さい単語数であれば有効である。

[0078] 実施の形態4.

上述の例では、インターネット上に提供されるハイパーテキストシステムを情報源としてWeb検索を行い、検索結果としてのスニップ及び関連文書を取得した。つまり、ワールドワイドウェブ(WWW)を検索対象とした。

[0079] しかし本発明は、他のデータベースを検索対象とする検索を行う場合にも有効である。他のデータベースで得られる要約や文書を前述のスニップ又は関連文書に置き換えて処理することにより、有効な応答が得られる。

[0080] 実施の形態5.

前述の相似形式質問文抽出処理(図12)では、各位置の単語の一致数を相似度としたが、他の基準により質問文形式要部同士の相似度を算出してもよい。

[0081] 例えば、質問文形式要部同士の単語列としての編集距離を算出し、編集距離を相似度として用いることも有効である。また、疑問詞より前の単語列同士の編集距離を算出し、更に疑問詞より後の単語列同士の編集距離を算出し、両編集距離の和を相似度とすることも有効である。

[0082] 尚、編集距離は、文字列同士がどの程度異なっているかを示す値であり、文字の削除、挿入、置換によって、一方の文字列を他方の文字列に変形するのに要する最小の手順回数として算出される。この例では、文字列に代えて、上述の形式化された単語列を扱うことにより編集距離を算出することができる。

[0083] 実施の形態6.

前述の相似形式質問文抽出処理(図12)では、各位置の単語の一致数を相似度としたが、各位置により重み付けを行うことも有効である。

[0084] 例えば、疑問詞からの距離に応じて、距離が小さい位置の一致の場合には大きい値を加算し、距離が大きい位置の一致の場合には小さい値を加算することにより、疑問詞近辺を重視する相似度を得ることもできる。

[0085] 逆に、距離が小さい位置の一致の場合には小さい値を加算し、距離が大きい位置の一致の場合には大きい値を加算することにより、疑問詞近辺を軽視する相似度を得ることもできる。

[0086] 実施の形態7.

また、相似度の算出の際に、単語の種類によって重み付けを行ってもよい。例えば、図7のS701からS704で判定した単語の種類毎に重みを設定し、単語が一致した場合に、その重みを加算することにより相似度を求めることが有効である。質問の焦点となりやすい単語に対して大きい重みを設定することなどが考えられる。

[0087] 実施の形態8.

上述の実施の形態では、応答形式要素相関度算出処理(図13)においてカイ二乗検定により応答形式要素の質問文に対する質問形式相関度を算出したが、他の基準により質問形式相関度を算出することもできる。

[0088] 参考応答例を全体の集合として、集合Aとして参考応答例形式記憶部104に含まれる参考応答ID群と、集合Bとして応答形式要素含回答文集合記憶部910に含まれる参考応答ID数群の要素数が計数可能であるので、例えば、ダイス係数を質問形式相関度として用いることもできる。

[0089] その場合には、集合Aの要素数として相似形式質問文集合記憶部908に含まれる参考応答ID数を計数し分母第1項値とし、集合Bの要素数として応答形式要素含回答文集合記憶部910に含まれる参考応答ID数を計数し分母第2項値とし、分母第1項値と分母第2項値を合計して、分母値を求める。また、集合Aと集合Bの積集合の要素数として相似形式質問文集合記憶部908と参考応答例形式記憶部104に共に含まれる参考応答ID数を計数し、それに2を乗じて分子値とする。そして、分子値を分母値で割ることによりダイス係数を算出し、それを質問形式相関度とする。

[0090] 実施の形態9.

また、相互情報量を質問形式相関度とすることもできる。

[0091] その場合には、集合Aの要素数として相似形式質問文集合記憶部908に含まれる参考応答ID数を計数し分母第1項値とし、集合Bの要素数として応答形式要素含回答文集合記憶部910に含まれる参考応答ID数を計数し分母第2項値とし、分母第1項値と分母第2項値を積算して、分母値を求める。また、集合Aと集合Bの積集合の要素数として相似形式質問文集合記憶部908と参考応答例形式記憶部104に共に含まれる参考応答ID数を計数し、それに全参考応答例数を乗じて分子値とする。そして、分子値を分母値で割り、その商に対する底を2とする対数を算出して、相互情

報量を得る。そして、それを質問形式相関度とする。

[0092] いずれの質問形式相関度の算出方法も、応答形式要素を含む参考回答文形式化単語列と相似形式質問文が組み合せられる確率に基づき、当該応答形式要素が相似形式質問文に形式として関連する程度を算出している。

[0093] 実施の形態10.

Web検索エンジンを用いて質問応答システムを実現することができる。図26は、Web検索エンジンを用いる質問応答システムの構成を示す図である。質問応答システムは、参考文形式化部2601、参考文記憶部2602、入力質問文形式化部2603、参考回答文抽出部2604、Web検索要求部2605、記述スタイル評価部2606、関連性評価部2607、スコア処理部2608、及び回答文出力部2609を有している。

[0094] 図27は、Web検索エンジンを用いる質問応答システムの処理フローを示す図である。参考文形式化部2601による参考文形式化処理(S2701)では、参考質問文と参考回答文の対を参考文とし、参考文のうち少なくとも参考質問文に対して記述スタイルを一般化する形式化処理を行う。参考文記憶部2602は、参考文形式化部2601において形式化された形式化参考文を記憶する。入力質問文形式化部2603による入力質問文形式化処理(S2702)では、入力質問文の記述スタイルを一般化する形式化処理を行う。参考回答文抽出部2604による参考回答文抽出処理(S2703)では、入力質問文形式化部2603において形式化された形式化入力質問文と類似する形式を有する形式化参考文を探索し、この形式化参考文に含まれる参考回答文を参考文記憶部2602から抽出する。

[0095] Web検索要求部2605によるWeb検索要求処理(S2704)では、入力質問文を条件としてWeb検索エンジンにWeb上の文書の検索を要求し、結果として検索Web文書を得る。記述スタイル評価部2606による記述スタイル評価処理(S2705)では、参考回答文と検索Web文書の間の記述スタイルの適合性を評価する。関連性評価部2607による関連性評価処理(S2706)では、検索Web文書と入力質問文の間の内容の関連性を評価する。

[0096] スコア処理部2608によるスコア処理(S2707)では、記述スタイル評価部により参考回答文と記述スタイルの適合性があると評価され、かつ、関連性評価部により前記

入力質問文の内容と関連があると評価された検索Web文書に対してスコア付けを行う。そして、回答文出力部2609による回答文出力処理(S2708)では、2608スコア処理部から入力質問文に対する回答文を得て、出力する。

[0097] 質問応答システムは、コンピュータであり、各要素はプログラムにより処理を実行することができる。また、プログラムを記憶媒体に記憶させ、記憶媒体からコンピュータに読み取られるようにすることができる。

[0098] 質問応答システムのハードウェアの構成について説明する。図28は、質問応答システムのハードウェアの構成を示す図である。バスに、演算装置2801、データ記憶装置2802、メモリ2803、通信インターフェース2804、データ入力装置2805、データ出力装置2806が接続されている。データ記憶装置2802は、例えばROM(Read Only Memory)やハードディスクである。メモリ2803は、通常RAM(Random Access Memory)である。プログラムは、通常データ記憶装置2802に記憶されており、メモリ2803にロードされた状態で、順次演算装置2801に読み込まれ処理を行う。通信インターフェース2804は、ネットワークを介した通信に用いる。データ入力装置2805は、データの入力に用いる。データ出力装置2806は、データの出力に用いる。

図面の簡単な説明

[0099] [図1]参考応答例準備処理に係る構成を示す図である。

[図2]参考応答例準備処理フローを示す図である。

[図3]参考応答例記憶部の構成例を示す図である。

[図4]参考応答例形式化処理フローを示す図である。

[図5]参考応答例形式記憶部の構成例を示す図である。

[図6]質問文形式化処理／回答文形式化処理フローを示す図である。

[図7]質問応答特性判定処理フローを示す図である。

[図8]質問応答処理フローを示す図である。

[図9]質問文入力から応答形式要素相関度計算までの処理に係る構成を示す図である。

[図10]質問文と参考質問文の比較例を示す図である。

[図11]質問文形式要部抽出処理フローを示す図である。

[図12]相似形式質問文抽出処理フローを示す図である。

[図13]応答形式要素相関度算出処理フローを示す図である。

[図14]応答形式要素相関度テーブルを示す図である。

[図15]関連語生成から応答文出力までの処理に係る構成を示す図である。

[図16]関連語生成処理フローを示す図である。

[図17]関連語テーブルの構成例を示す図である。

[図18]関連文書検索処理フローを示す図である。

[図19]文スコア算出処理フローを示す図である。

[図20]内容評価項算出処理フローを示す図である。

[図21]形式評価項算出処理フローを示す図である。

[図22]文スコアの分布例を示す図である。

[図23]解候補抽出処理フローを示す図である。

[図24]応答文出力処理フローを示す図である。

[図25]正解と不正解の例を示す図である。

[図26]Web検索エンジンを用いる質問応答システムの構成を示す図である。

[図27]Web検索エンジンを用いる質問応答システムの処理フローを示す図である。

[図28]質問応答システムのハードウェアの構成を示す図である。

符号の説明

[0100] 101 参考応答例生成部、102 参考応答例記憶部、103 参考応答例形式化部、104 参考応答例形式記憶部、901 質問文入力部、902 質問文記憶部、903 質問文形式化部、904 質問文形式記憶部、905 質問文形式要部抽出部、906 質問文形式要部記憶部、907 相似形式質問文抽出部、908 相似形式質問文集合記憶部、909 応答形式要素相関度算出部、910 応答形式要素含回答文集合記憶部、911 応答形式要素相関度テーブル、1501 関連語生成部、1502 関連語テーブル、1503 関連文書検索部、1504 関連文書記憶部、1505 文スコア算出部、1506 文スコアテーブル、1507 解候補抽出部、1508 解候補記憶部、1509 応答文出力部、2601 参考文形式化部、2602 参考文記憶部、2603 入力

質問文形式化部、2604　参考回答文抽出部、2605　Web検索要求部、2606　記述スタイル評価部、2607　関連性評価部、2608　スコア処理部、2609　回答文出力部。

請求の範囲

[1] 質問文を入力し、検索対象である文書群から質問文の解に適する文を抽出して、応答文として出力する質問応答システムであって、以下の要素を有することを特徴とする質問応答システム

(1)質問文を入力する質問文入力部

(2)文を単語に分割し、単語毎に品詞種別と表層表現を解析し、各単語が機能語である場合、疑問詞である場合、及び質問の焦点になりやすい所定語である場合に、当該単語を質問応答形式に係る単語であると判定し、それ以外の場合に、当該単語を質問応答内容に係る単語であると判定し、質問応答形式に係る単語は品詞種別に変換し、質問応答内容に係る単語は表層表現に変換し、変換した品詞種別あるいは表層表現を単位とした形式化単語列とする形式化処理により、入力された質問文を質問文形式化単語列に変換する質問文形式化部

(3)疑問詞を所定位置に含む所定単語数の形式化単語列を抜き出す質問文形式要部抽出処理により、質問文形式化単語列から質問文形式要部を抽出する質問文形式要部抽出部

(4)参考質問文と参考回答文の対からなる参考応答例を複数記憶する参考応答例記憶部

(5)参考応答例記憶部に含まれる各参考応答例について、参考質問文を前記形式化処理により参考質問文形式化単語列に変換し、更に参考回答文を前記形式化処理により参考回答文形式化単語列に変換する参考応答例形式化部

(6)変換された参考質問文形式化単語列と参考回答文形式化単語列の対を、参考応答IDに対応付けて複数記憶する応答例形式記憶部

(7)応答例形式記憶部に含まれる各参考質問文形式化単語列について、前記質問文形式要部抽出処理により参考質問文形式要部を抽出し、前記質問文形式要部と比較し、比較結果が同一又は類似の場合に、当該参考質問文形式要部が抽出された参考質問文形式化単語列の参考応答IDを、入力された質問文に形式が相似する相似形式質問文に係る参考応答IDとして特定する相似形式質問文抽出部

(8)特定された相似形式質問文に係る参考応答ID群を、相似形式質問文集合とし

て記憶する相似形式質問文集合記憶部

(9)各相似形式質問文に係る参考応答IDに対応する参考回答文形式化単語列を応答例形式記憶部から取得し、前記形式化単語列の単語数よりも少ない所定単語数の形式化単語列である応答形式要素を、取得した参考回答文形式化単語列から順に抽出し、各応答形式要素について、当該応答形式要素が応答例形式記憶部に含まれる各参考回答文形式化単語列に含まれるか検索し、当該応答形式要素が含まれる参考回答文形式化単語列に係る参考応答ID群を応答形式要素含回答文集合として記憶し、少なくとも相似形式質問文集合と応答形式要素含回答文集合の両方に含まれる参考応答ID群の数、相似形式質問文集合に含まれる参考応答ID群の数、及び応答形式要素含回答文集合に含まれる参考応答ID群の数を用いて、当該応答形式要素を含む参考回答文形式化単語列と相似形式質問文が組み合せられる確率に基づき、当該応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度を算出する応答形式要素相関度算出部

(10)応答形式要素毎に算出された質問形式相関度を記憶する応答形式要素相関度テーブル

(11)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から文書を検索し、検索した文書群に含まれる単語の出現頻度に基づいて、内容として質問文に関連する関連語を抽出するとともに当該関連語の関連度を算出する関連語生成部

(12)関連語毎に算出された関連度を記憶する関連語テーブル

(13)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から関連文書を検索する検索関連文書検索部

(14)検索された関連文書を、関連文書に含まれる関連文毎に文番号を対応付けて記憶する関連文書記憶部

(15)各関連文について、当該関連文を前記形式化処理により関連文形式化単語列に変換し、関連文形式化単語列から前記応答形式要素を順に抽出し、各応答形式要素の質問形式相関度を応答形式要素相関度テーブルから取得し、更に当該関連文に含まれる各単語の関連語としての関連度を関連語テーブルから取得し、取得し

た各応答形式要素の質問形式相関度及び各単語の関連度に基づいて、質問文に対する解としての適性を示す文スコアを算出する文スコア算出部

(16)関連文毎の文スコアを文番号に対応付けて記憶する文スコアテーブル

(17)高い適性を示す文スコアの関連文の文番号を解候補として抽出する解候補抽出部

(18)解候補の文番号により特定される関連文を応答文として出力する応答文出力部。

[2] 前記質問の焦点になりやすい所定語として、少なくとも「理由」、「方法」、「意味」、又は「違い」の何れかを用いることを特徴とする請求項1記載の質問応答システム。

[3] 前記形式化処理は、各単語が参考応答例の中で出現頻度が高い所定の動詞と形容詞である場合にも、当該単語を質問応答形式に係る単語であると判定することを特徴とする請求項1記載の質問応答システム。

[4] 前記質問文形式要部抽出処理により抜き出される形式化単語列は、疑問詞を中心として前後3つ単語を含む合計7つの単語に係る形式化単語列であることを特徴とする請求項1記載の質問応答システム。

[5] 前記相似形式質問文抽出部は、参考質問文形式要部と質問文形式要部に含まれる疑問詞が一致する場合に限り、類似と判定することを特徴とする請求項1記載の質問応答システム。

[6] 前記応答形式要素相関度算出部は、応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度として、カイ二乗値の平方根を用いることを特徴とする請求項1記載の質問応答システム。

[7] 前記応答形式要素相関度算出部は、応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度として、ダイス係数を用いることを特徴とする請求項1記載の質問応答システム。

[8] 前記応答形式要素相関度算出部は、応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度として、相互情報量を用いることを特徴とする請求項1記載の質問応答システム。

[9] 前記解候補抽出部は、関連文書に含まれる関連文の順に連続する文スコアについ

て、極大値を示す文スコアの関連文の文番号を解候補とすることを特徴とする請求項1記載の質問応答システム。

[10] 前記解候補抽出部は、前記極大値の所定割合を超える前後の文スコアの関連文の文番号も解候補に含めることを特徴とする請求項9記載の質問応答システム。

[11] 質問文を入力し、検索対象である文書群から質問文の解に適する文を抽出して、応答文として出力する質問応答システムであって、

参考質問文と参考回答文の対からなる参考応答例を複数記憶する参考応答例記憶部と、

参考質問文形式化単語列と参考回答文形式化単語列の対を、参考応答IDに対応付けて複数記憶するための応答例形式記憶部と、

相似形式質問文に係る参考応答ID群を、相似形式質問文集合として記憶するための相似形式質問文集合記憶部と、

応答形式要素毎に算出された質問形式相関度を記憶するための応答形式要素相関度テーブルと、

関連語毎に算出された関連度を記憶するための関連語テーブルと、

関連文書を、関連文書に含まれる関連文毎に文番号を対応付けて記憶するための関連文書記憶部と、

関連文毎の文スコアを文番号に対応付けて記憶するための文スコアテーブルと、

を有する質問応答システムとなるコンピュータに、以下の手順を実行させるためのプログラム

(1)質問文を入力する質問文入力手順

(2)文を単語に分割し、単語毎に品詞種別と表層表現を解析し、各単語が機能語である場合、疑問詞である場合、及び質問の焦点になりやすい所定語である場合に、当該単語を質問応答形式に係る単語であると判定し、それ以外の場合に、当該単語を質問応答内容に係る単語であると判定し、質問応答形式に係る単語は品詞種別に変換し、質問応答内容に係る単語は表層表現に変換し、変換した品詞種別あるいは表層表現を単位とした形式化単語列とする形式化処理により、入力された質問文を質問文形式化単語列に変換する質問文形式化手順

(3)疑問詞を所定位置に含む所定単語数の形式化単語列を抜き出す質問文形式要部抽出処理により、質問文形式化単語列から質問文形式要部を抽出する質問文形式要部抽出手順

(4)参考応答例記憶部に含まれる各参考応答例について、参考質問文を前記形式化処理により参考質問文形式化単語列に変換し、更に参考回答文を前記形式化処理により参考回答文形式化単語列に変換する参考応答例形式化手順

(5)応答例形式記憶部に含まれる各参考質問文形式化単語列について、前記質問文形式要部抽出処理により参考質問文形式要部を抽出し、前記質問文形式要部と比較し、比較結果が同一又は類似の場合に、当該参考質問文形式要部が抽出された参考質問文形式化単語列の参考応答IDを、入力された質問文に形式が相似する相似形式質問文に係る参考応答IDとして特定する相似形式質問文抽出手順

(6)各相似形式質問文に係る参考応答IDに対応する参考回答文形式化単語列を応答例形式記憶部から取得し、前記形式化単語列の単語数よりも少ない所定単語数の形式化単語列である応答形式要素を、取得した参考回答文形式化単語列から順に抽出し、各応答形式要素について、当該応答形式要素が応答例形式記憶部に含まれる各参考回答文形式化単語列に含まれるか検索し、当該応答形式要素が含まれる参考回答文形式化単語列に係る参考応答ID群を応答形式要素含回答文集合として記憶し、少なくとも相似形式質問文集合と応答形式要素含回答文集合の両方に含まれる参考応答ID群の数、相似形式質問文集合に含まれる参考応答ID群の数、及び応答形式要素含回答文集合に含まれる参考応答ID群の数を用いて、当該応答形式要素を含む参考回答文形式化単語列と相似形式質問文が組み合せられる確率に基づき、当該応答形式要素が相似形式質問文に形式として関連する程度を示す質問形式相関度を算出する応答形式要素相関度算出手順

(7)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象の文書群から文書を検索し、検索した文書群に含まれる単語の出現頻度に基づいて、内容として質問文に関連する関連語を抽出するとともに当該関連語の関連度を算出する関連語生成手順

(8)質問文から内容語であるキーワードを抽出し、キーワードを条件として検索対象

の文書群から関連文書を検索する検索関連文書検索手順

(9)各関連文について、当該関連文を前記形式化処理により関連文形式化単語列に変換し、関連文形式化単語列から前記応答形式要素を順に抽出し、各応答形式要素の質問形式相関度を応答形式要素相関度テーブルから取得し、更に当該関連文に含まれる各単語の関連語としての関連度を関連語テーブルから取得し、取得した各応答形式要素の質問形式相関度及び各単語の関連度に基づいて、質問文に対する解としての適性を示す文スコアを算出する文スコア算出手順

(10)高い適性を示す文スコアの関連文の文番号を解候補として抽出する解候補抽出手順

(11)解候補の文番号により特定される関連文を応答文として出力する応答文出力手順。

[12] 参考質問文と参考回答文の対を参考文とし、該参考文のうち少なくとも参考質問文に対して記述スタイルを一般化する形式化処理を行う参考文形式化部と、

前記参考文形式化部において形式化された形式化参考文を記憶する参考文記憶部と、

入力質問文の記述スタイルを一般化する形式化処理を行う入力質問文形式化部と、

前記入力質問文形式化部において形式化された形式化入力質問文と類似する形式を有する前記形式化参考文を探索し、該形式化参考文に含まれる参考回答文を前記参考文記憶部から抽出する参考回答文抽出部と、

前記参考回答文と、前記入力質問文をWebサーチエンジンで検索した結果得られたWeb文書である検索Web文書との間の記述スタイルの適合性を評価する記述スタイル評価部と、

前記検索Web文書と、前記入力質問文との間の内容の関連性を評価する関連性評価部と、

前記記述スタイル評価部により前記参考回答文と記述スタイルの適合性があると評価され、かつ、前記関連性評価部により前記入力質問文の内容と関連があると評価された検索Web文書に対してスコア付け処理を行うスコア処理部と、

該スコアに基づいて、前記入力質問文に対する回答文を出力する回答文出力部を有することを特徴とする質問応答システム。

[13] 質問応答システムとなるコンピュータに、

参考質問文と参考回答文の対を参考文とし、該参考文のうち少なくとも参考質問文に対して記述スタイルを一般化する形式化処理を行う参考文形式化手順と、

前記参考文形式化手順において形式化された形式化参考文を記憶する参考文記憶手順と、

入力質問文の記述スタイルを一般化する形式化処理を行う入力質問文形式化手順と、

前記入力質問文形式化手順において形式化された形式化入力質問文と類似する形式を有する前記形式化参考文を探索し、該形式化参考文に含まれる参考回答文を抽出する参考回答文抽出手順と、

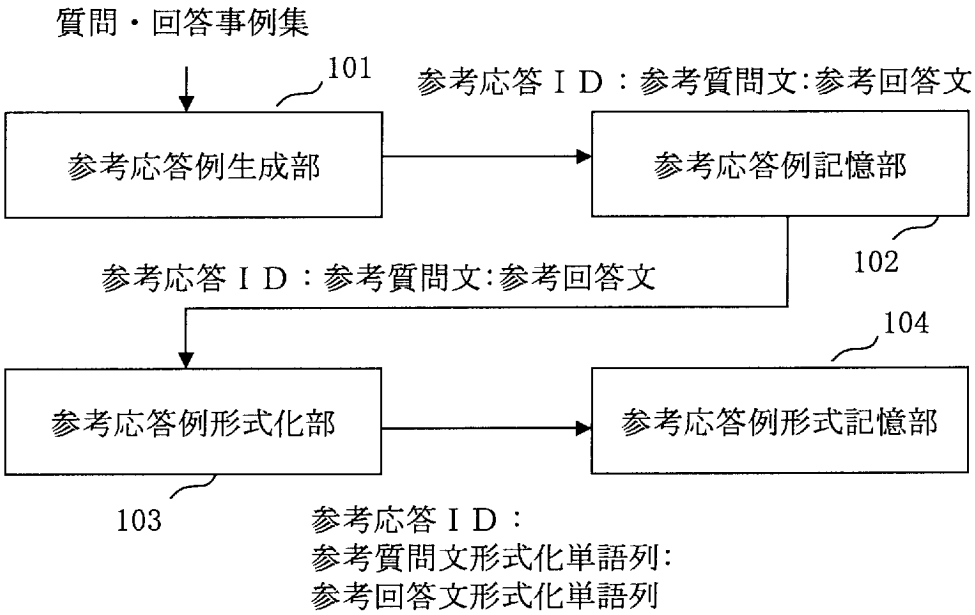
前記参考回答文と、前記入力質問文をWebサーチエンジンで検索した結果得られたWeb文書である検索Web文書との間の記述スタイルの適合性を評価する記述スタイル評価手順と、

前記検索Web文書と、前記入力質問文との間の内容の関連性を評価する関連性評価手順と、

前記記述スタイル評価手順により前記参考回答文と記述スタイルの適合性があると評価され、かつ、前記関連性評価手順により前記入力質問文の内容と関連があると評価された検索Web文書に対してスコア付け処理を行うスコア処理手順と、

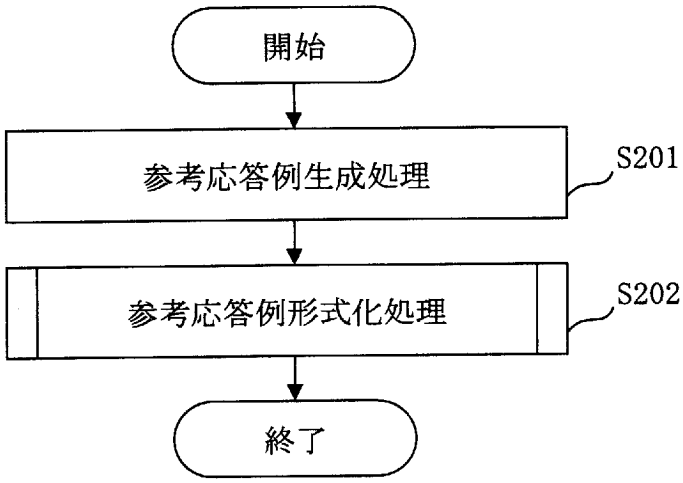
該スコアに基づいて、前記入力質問文に対する回答文を出力する回答文出力手順を実行させるためのプログラム。

[図1]



＜参考応答例準備処理に係る構成＞

[図2]



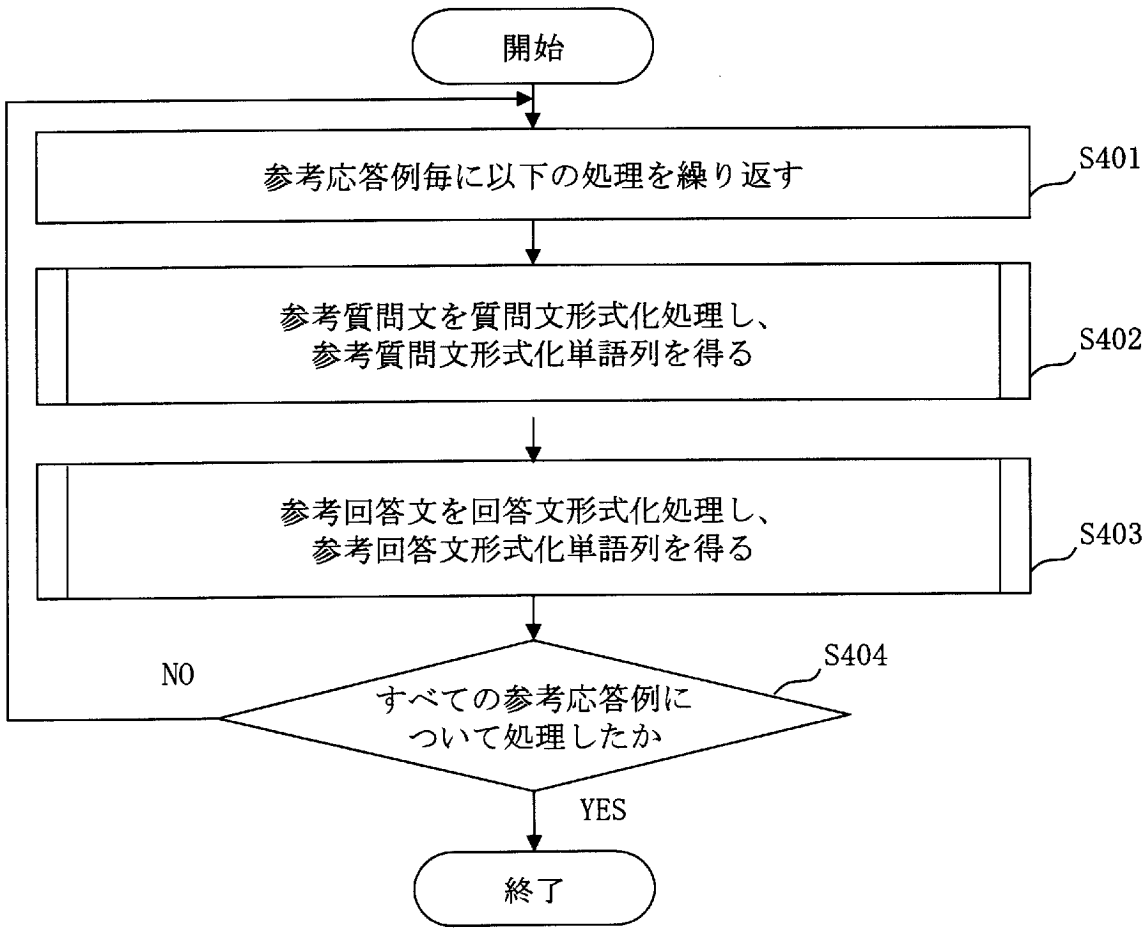
＜参考応答例準備処理フロー＞

[図3]

351 参考応答ＩＤ	352 参考質問文	353 参考回答文
⋮	⋮	⋮
ＱＡ０９９	～	～
ＱＡ１００	消費税込みの値段が 表示されるようになった 理由は何ですか。	表向きは値段を 分かり易くするため。
ＱＡ１０１	～	～
⋮	⋮	⋮

＜参考応答例記憶部＞

[図4]



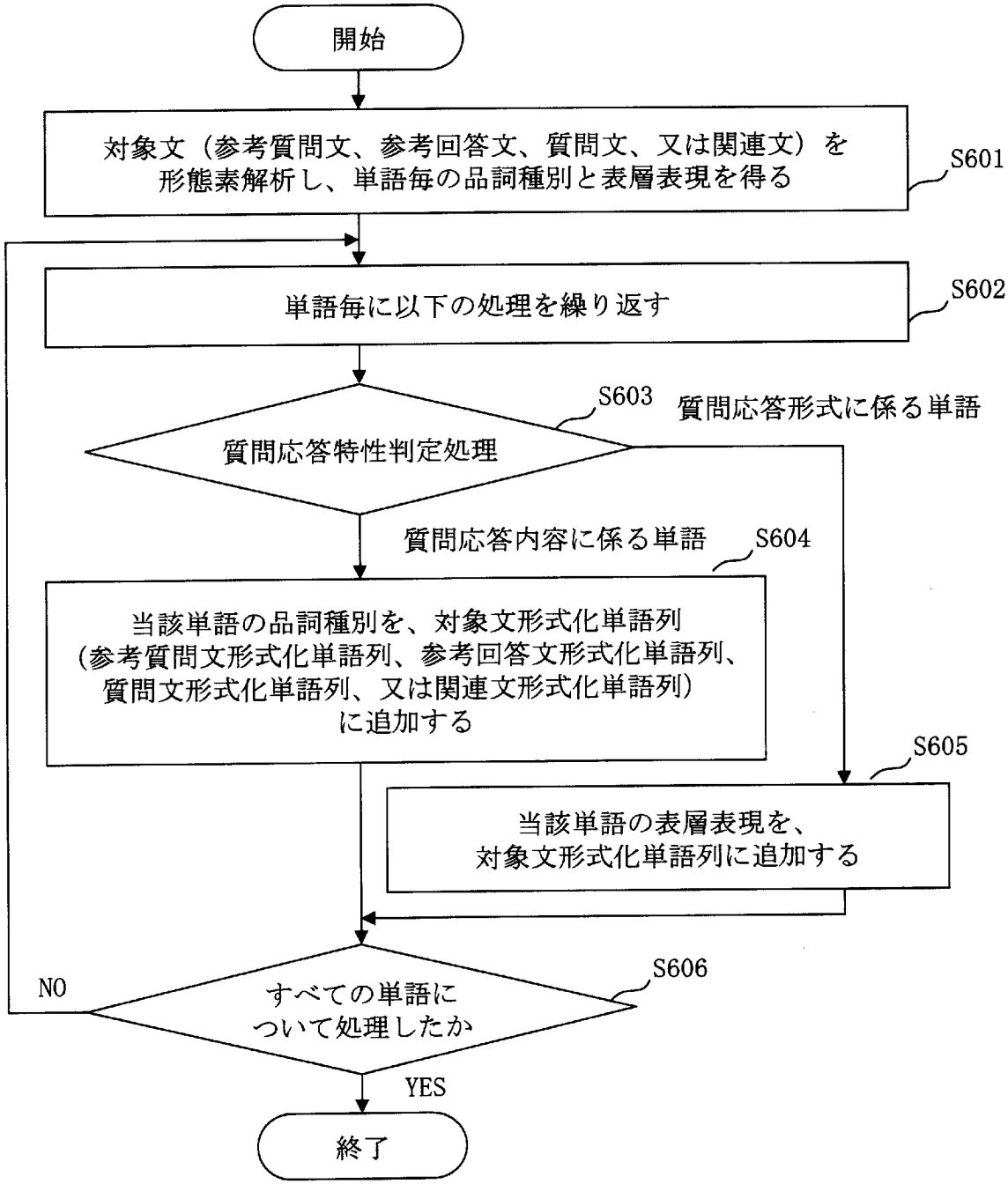
＜参考応答例形式化処理フロー＞

[図5]

551 参考応答ＩＤ	552 参考質問文形式化単語列	553 参考回答文形式化単語列
⋮	⋮	⋮
ＱＡ０９９	～	～
ＱＡ１００	＜名詞，サ変接続，＊，＊＞__ ＜名詞，一般，＊，＊＞__ノ__ ＜名詞，一般，＊，＊＞__ガ__ ＜名詞，サ変接続，＊，＊＞__ サ__レル__ヨウ__ニ__ナッ__ タ__リユウ__ハ__ナニ__デス__ カ__＜記号，句点，＊，＊＞	＜名詞，一般，＊，＊＞__ハ__ ＜名詞，一般，＊，＊＞__ヲ__ ＜動詞，自立，＊，＊＞__ ＜形容詞，自立，＊，＊＞__ スル__タメ__ ＜記号，句点，＊，＊＞
ＱＡ１０１	～	～
⋮	⋮	⋮

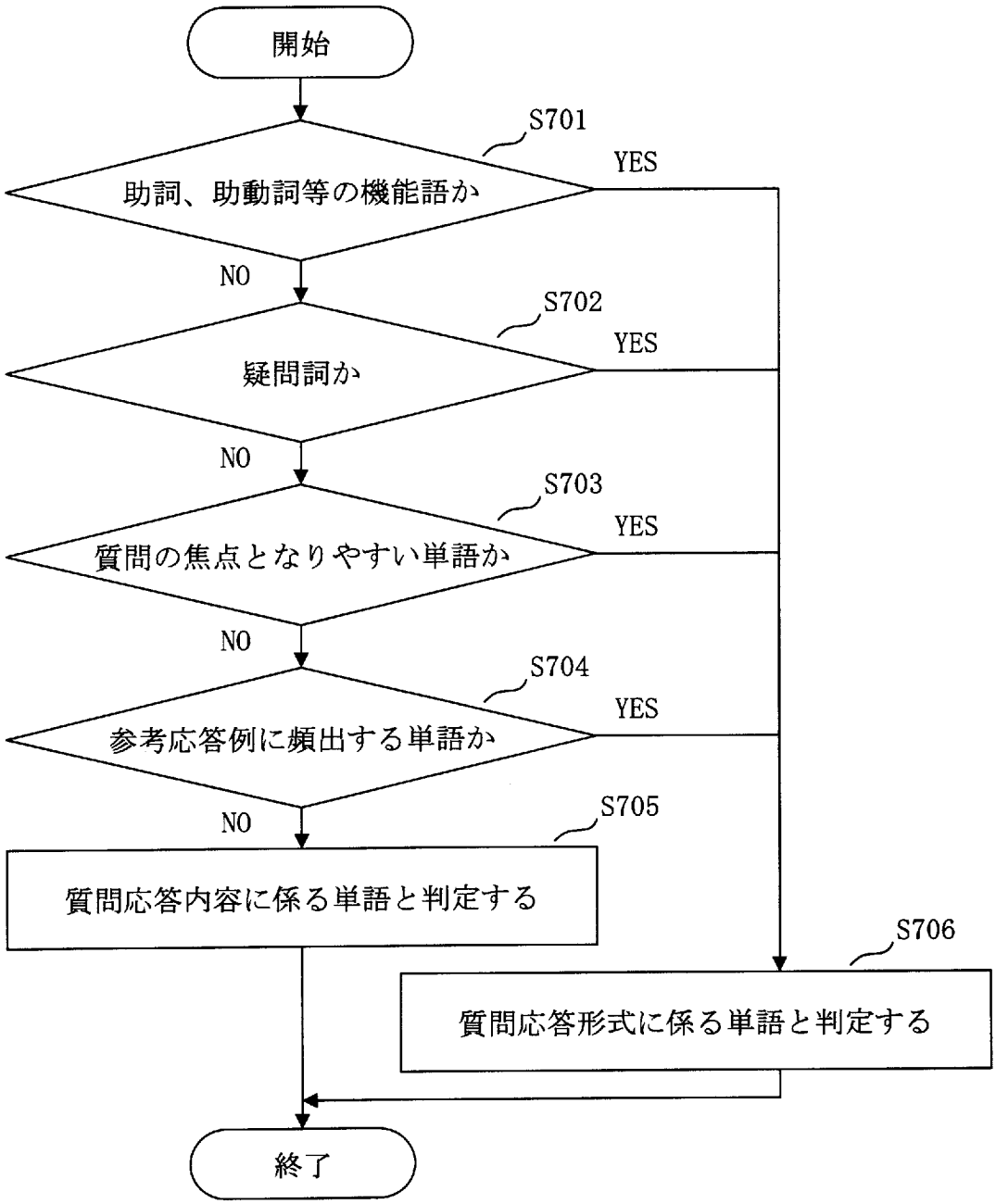
＜参考応答例形式記憶部＞

[図6]



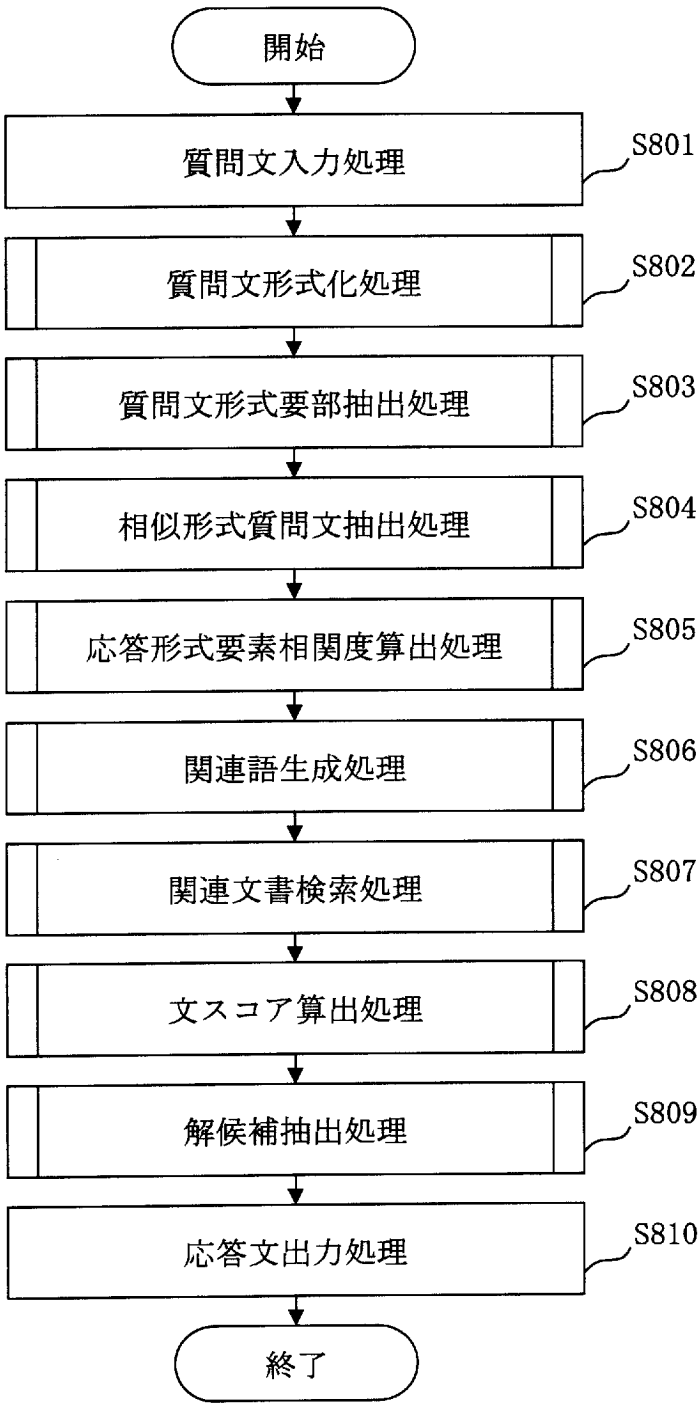
＜質問文形式化処理／回答文形式化処理フロー＞

[図7]



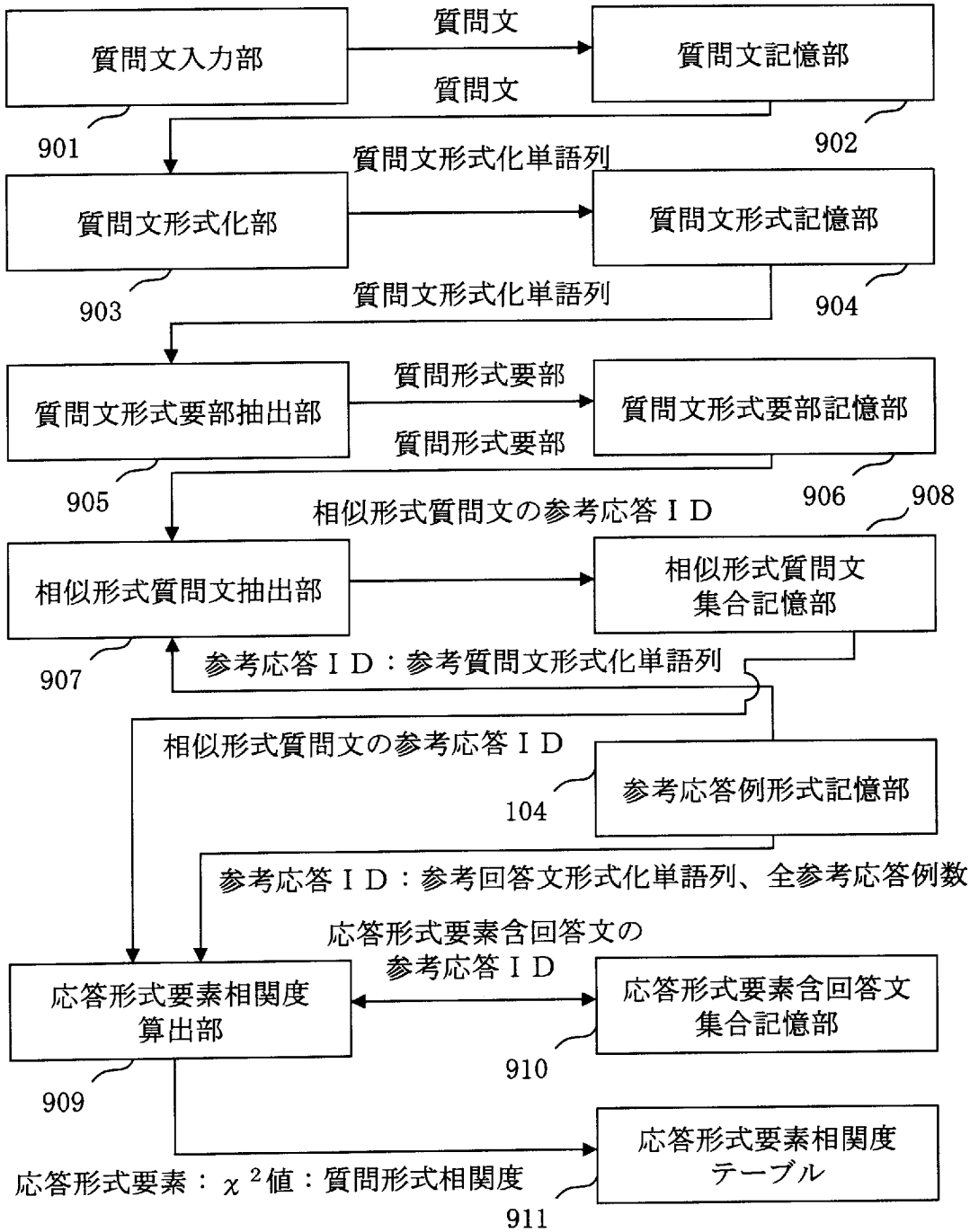
＜質問応答特性判定処理フロー＞

[図8]



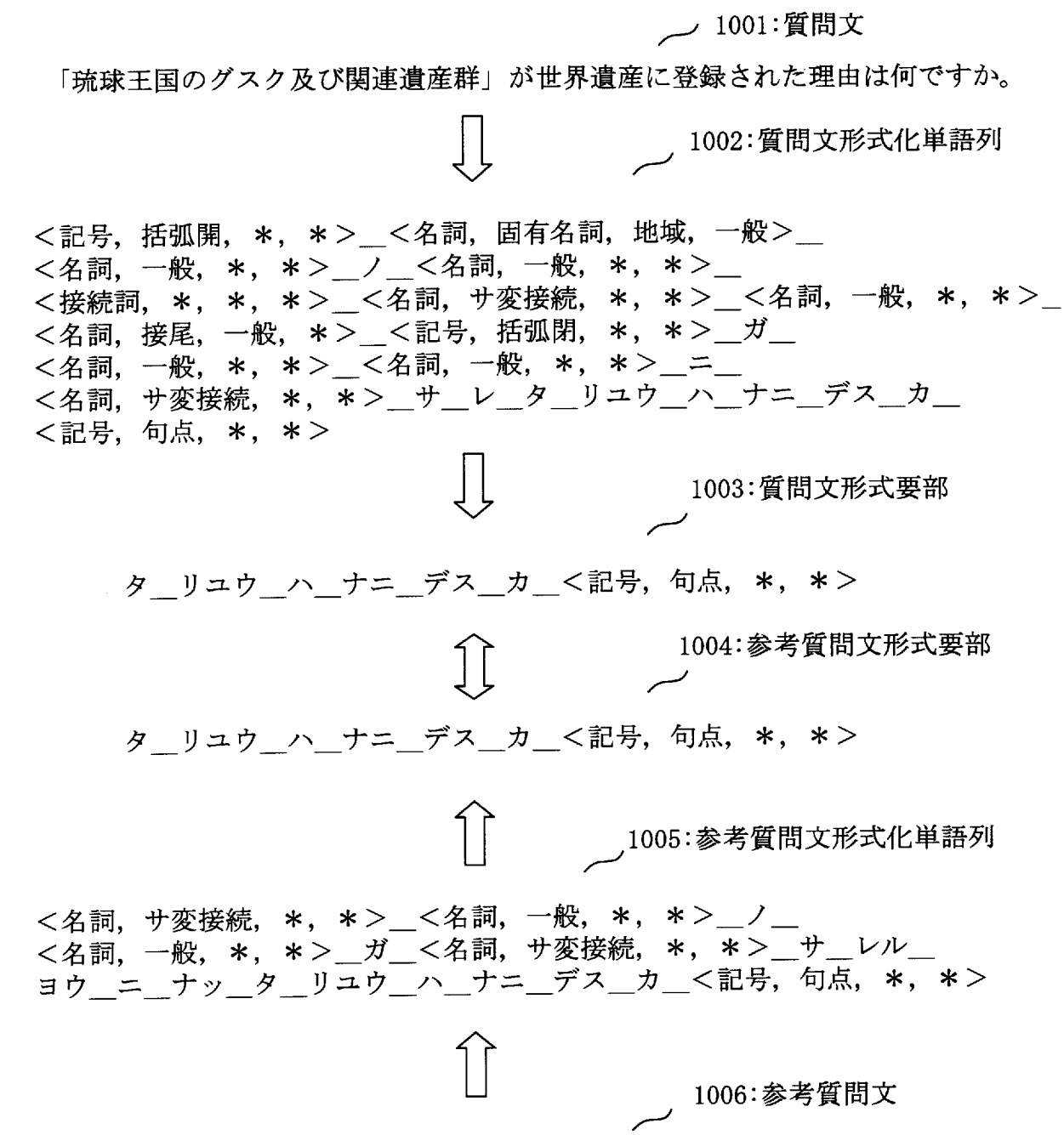
＜質問応答処理フロー＞

[図9]

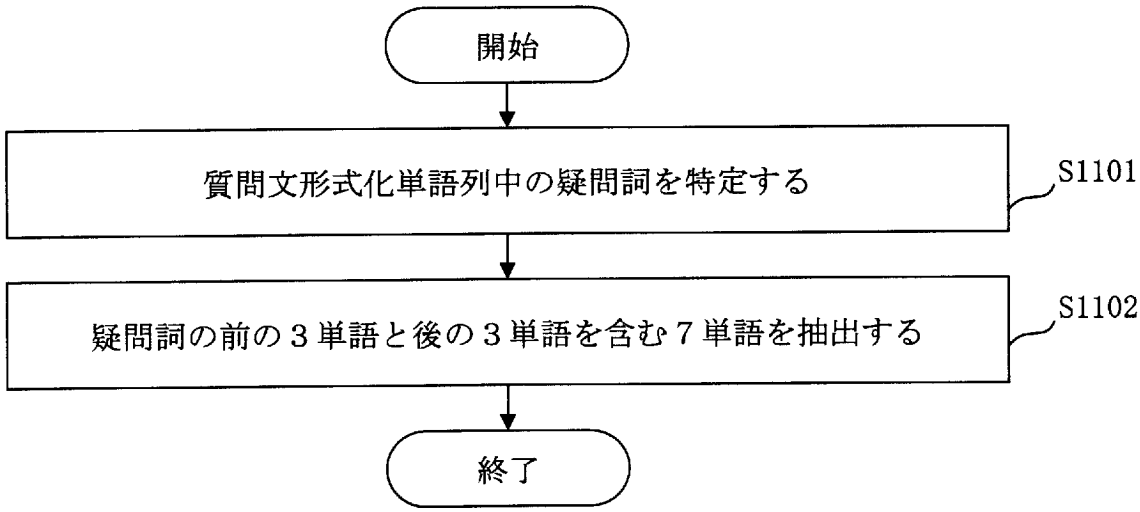


＜質問文入力から応答形式要素相関度算出までの処理に係る構成＞

[図10]

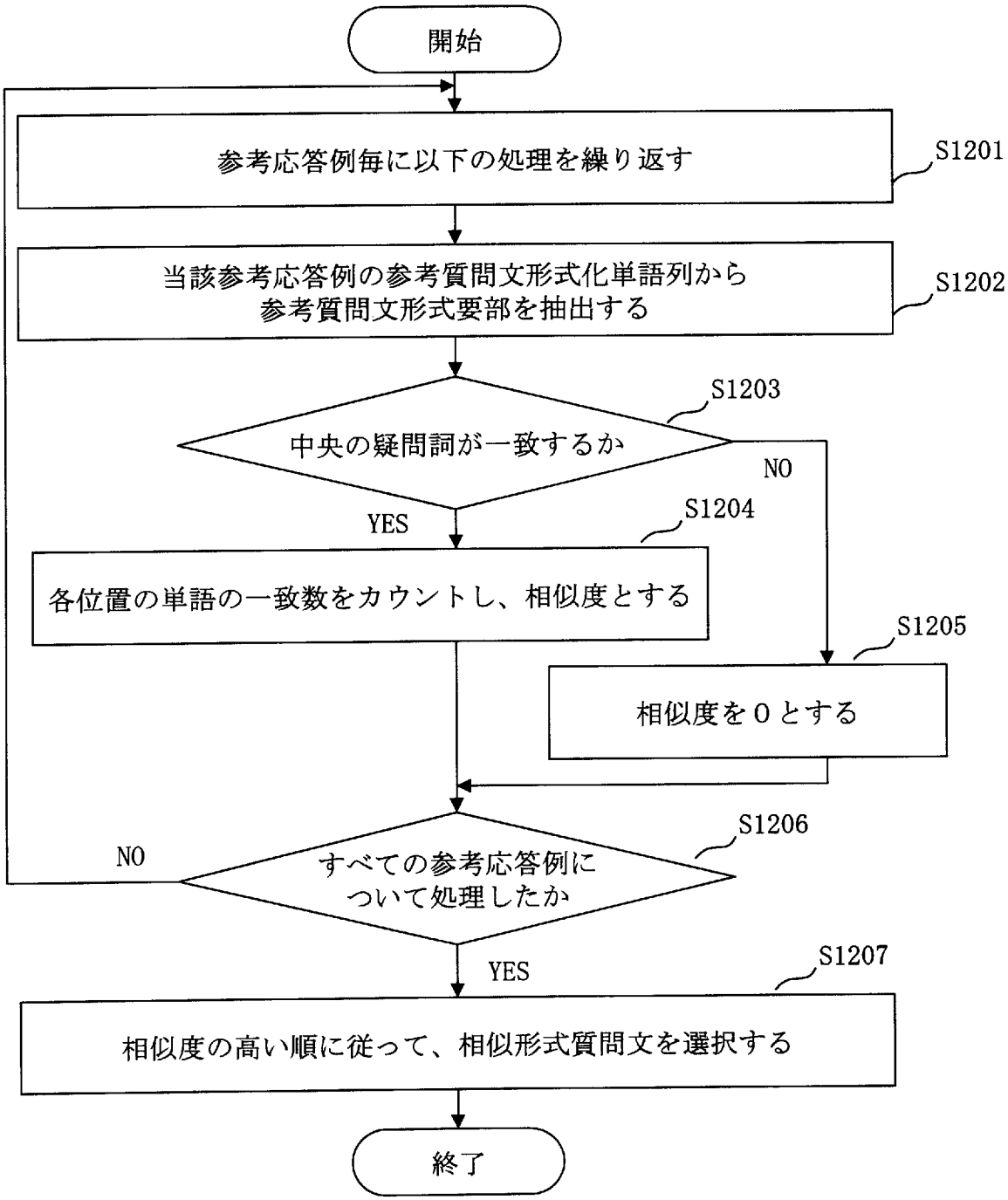

1001:質問文
「琉球王国のグスク及び関連遺産群」が世界遺産に登録された理由は何ですか。
1002:質問文形式化単語列
＜記号，括弧開，＊，＊＞_＜名詞，固有名詞，地域，一般＞_
＜名詞，一般，＊，＊＞_ノ_＜名詞，一般，＊，＊＞_
＜接続詞，＊，＊，＊＞_＜名詞，サ変接続，＊，＊＞_＜名詞，一般，＊，＊＞_
＜名詞，接尾，一般，＊＞_＜記号，括弧閉，＊，＊＞_ガ_
＜名詞，一般，＊，＊＞_＜名詞，一般，＊，＊＞_ニ_
＜名詞，サ変接続，＊，＊＞_サ_レ_タ_リユウ_ハ_ナニ_デス_カ_
＜記号，句点，＊，＊＞
1003:質問文形式要部
タ_リユウ_ハ_ナニ_デス_カ_＜記号，句点，＊，＊＞
1004:参考質問文形式要部
タ_リユウ_ハ_ナニ_デス_カ_＜記号，句点，＊，＊＞
1005:参考質問文形式化単語列
＜名詞，サ変接続，＊，＊＞_＜名詞，一般，＊，＊＞_ノ_
＜名詞，一般，＊，＊＞_ガ_＜名詞，サ変接続，＊，＊＞_サ_レル_
ヨウ_ニ_ナッ_タ_リユウ_ハ_ナニ_デス_カ_＜記号，句点，＊，＊＞
1006:参考質問文
消費税込みの値段が表示されるようになった理由は何ですか。

[図11]



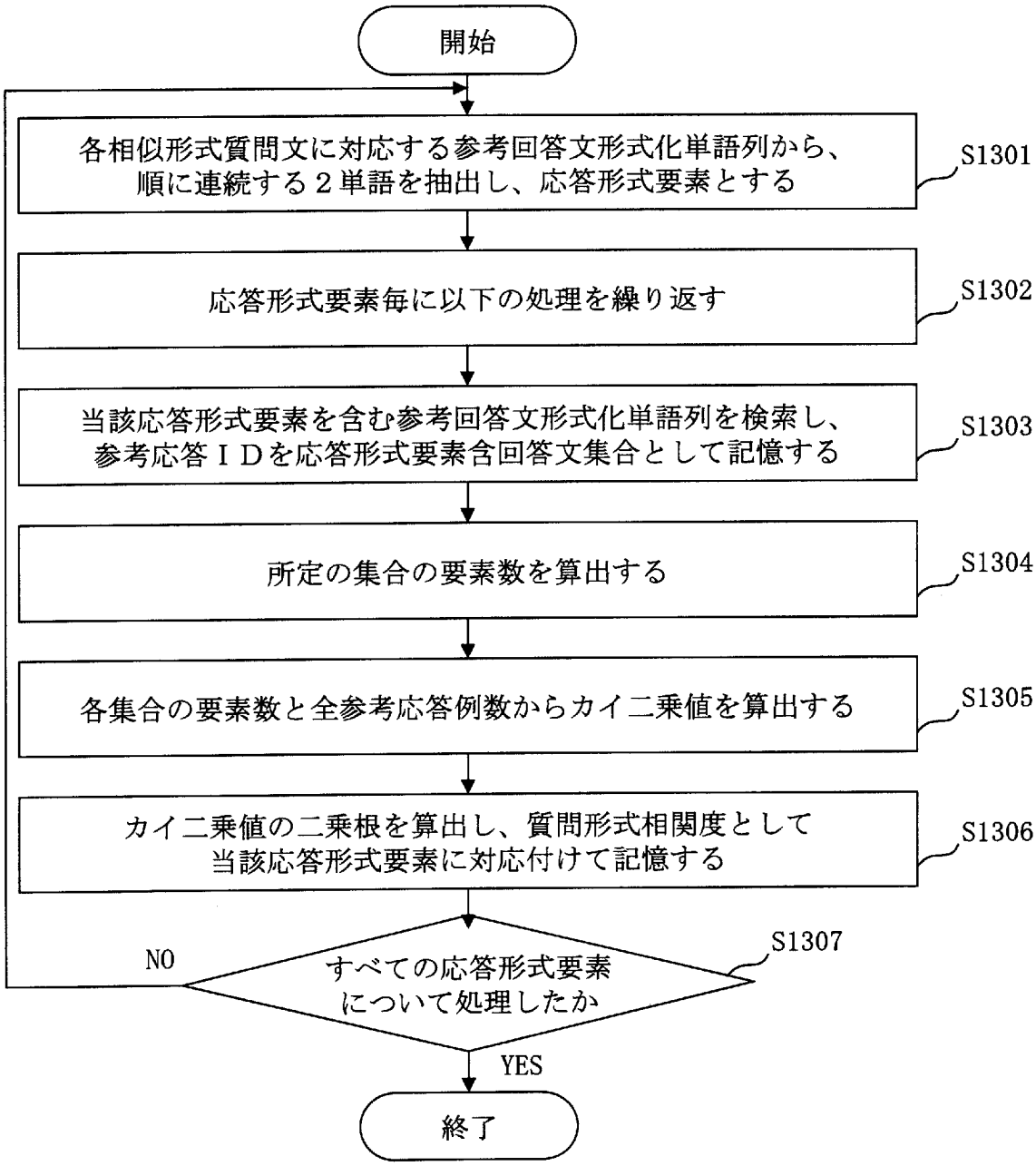
＜質問文形式要部抽出処理フロー＞

[図12]



＜相似形式質問文抽出処理フロー＞

[図13]



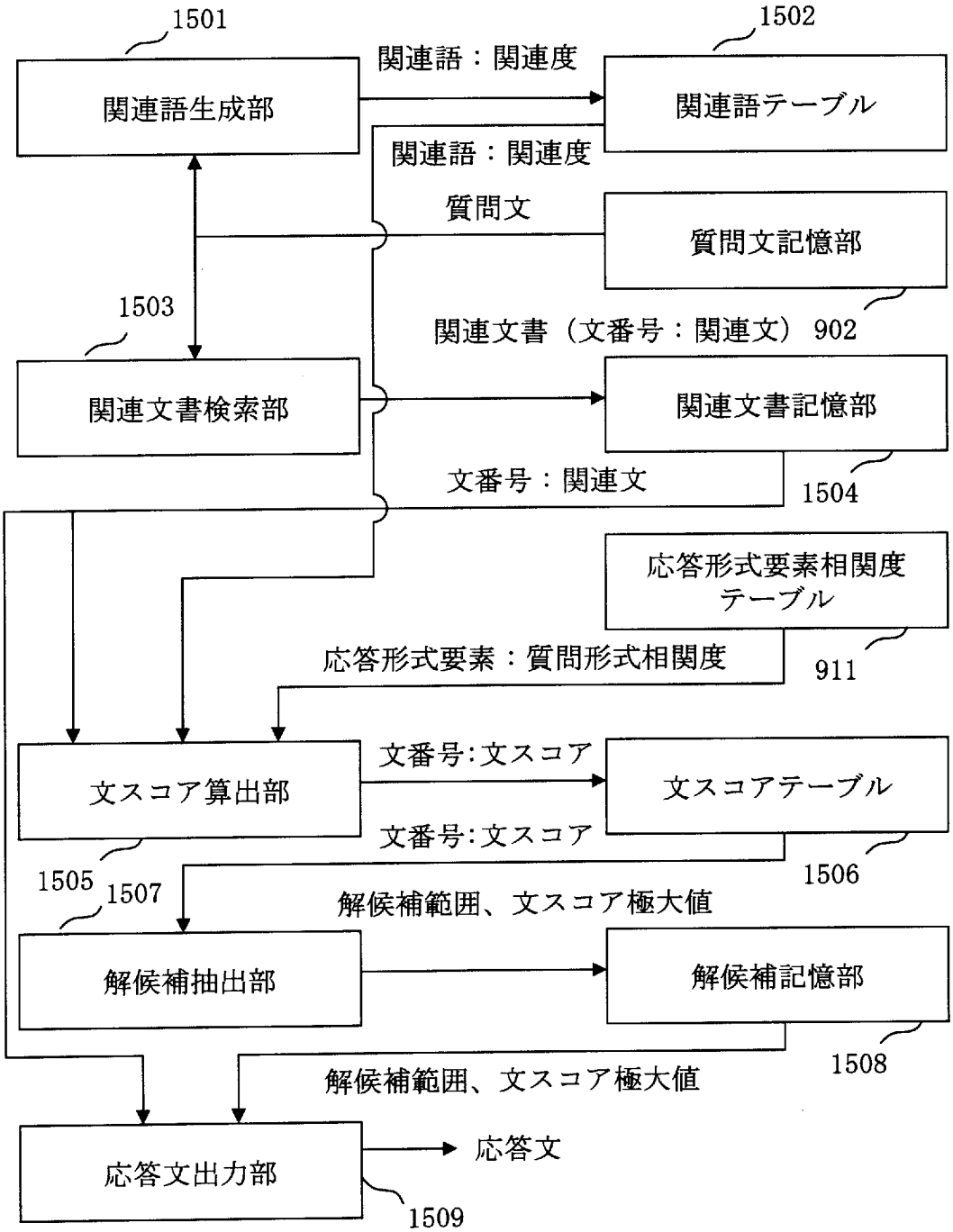
＜応答形式要素相関度算出処理フロー＞

[図14]

1451 応答形式要素	1452 χ^2値	1453 質問形式相関度
タ＿リユウ	７０５	（７０５）$^{1/2}$
タ＿カラ	５３１	（５３１）$^{1/2}$
リユウ＿ハ	２１９	（２１９）$^{1/2}$
⋮	⋮	⋮
カラ＿＜記号，句点，＊，＊＞	１１３	（１１３）$^{1/2}$
カラ＿デス	９８	（９８）$^{1/2}$
⋮	⋮	⋮
タメ＿＜記号，句点，＊，＊＞	４２	（４２）$^{1/2}$
タ＿ノデ	３４	（３４）$^{1/2}$
⋮	⋮	⋮
＜動詞，接尾，＊，＊＞＿＜記号，句点，＊，＊＞	１３	（１３）$^{1/2}$

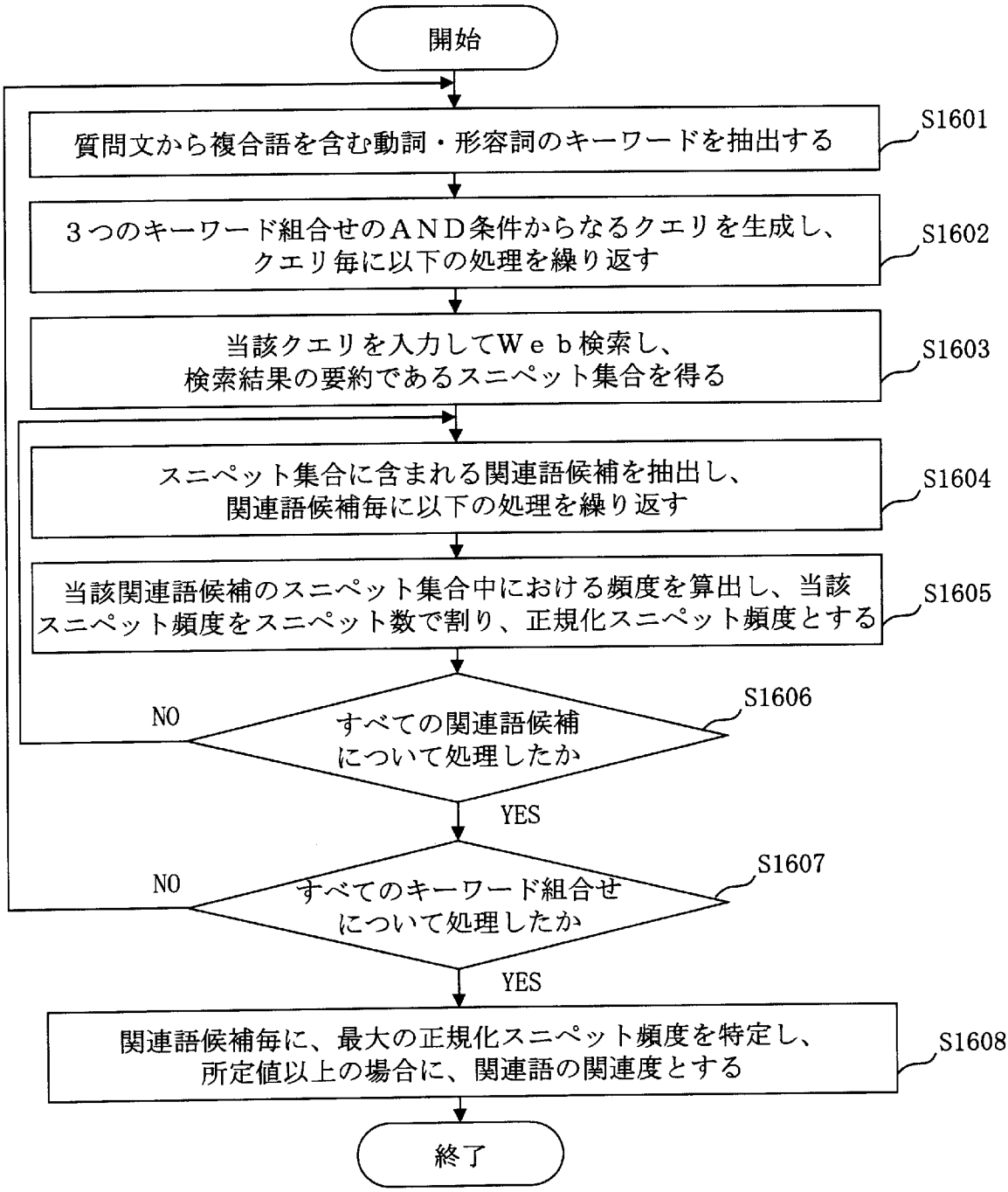
＜応答形式要素相関度テーブル＞

[図15]



＜関連語生成から応答文出力までの処理に係る構成＞

[図16]



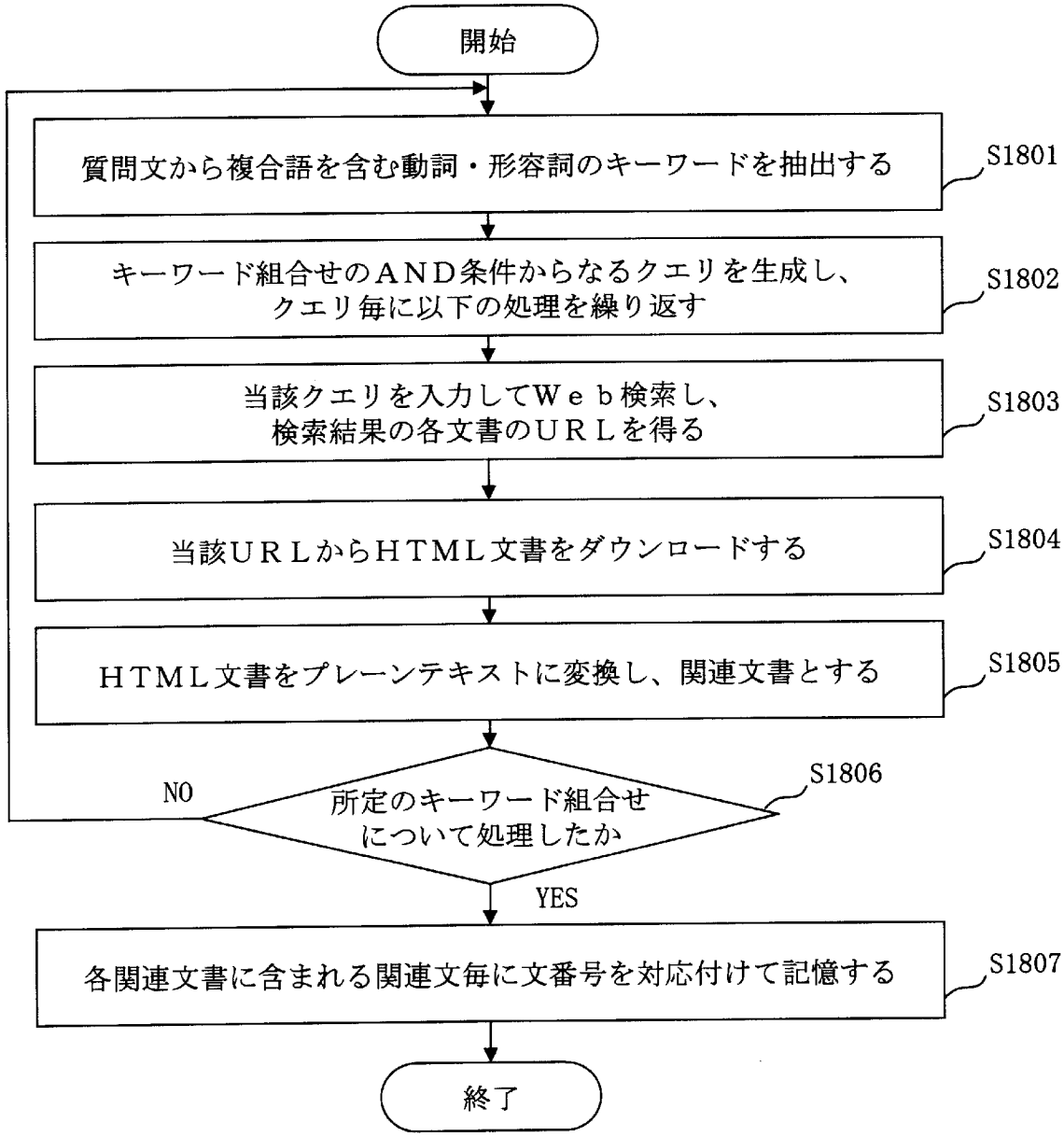
＜関連語生成処理フロー＞

[図17]

1751 関連語	1752 関連度
２０００	０．５０
沖縄	０．３８
文化	０．３７
自然	０．３４
日本	０．３１
城跡	０．２５
里	０．２５
首	０．２５
⋮	⋮

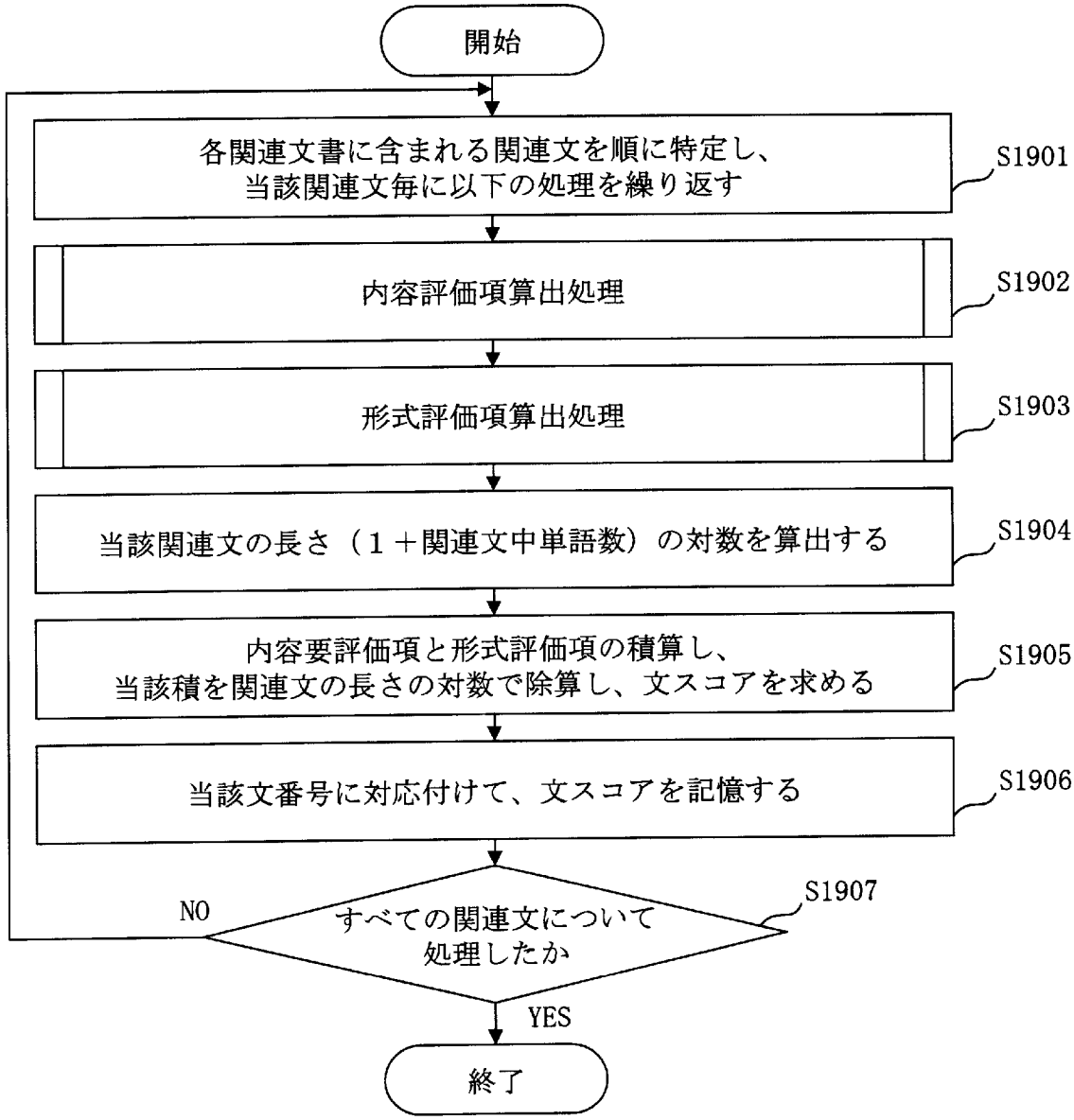
＜関連語テーブル＞

[図18]



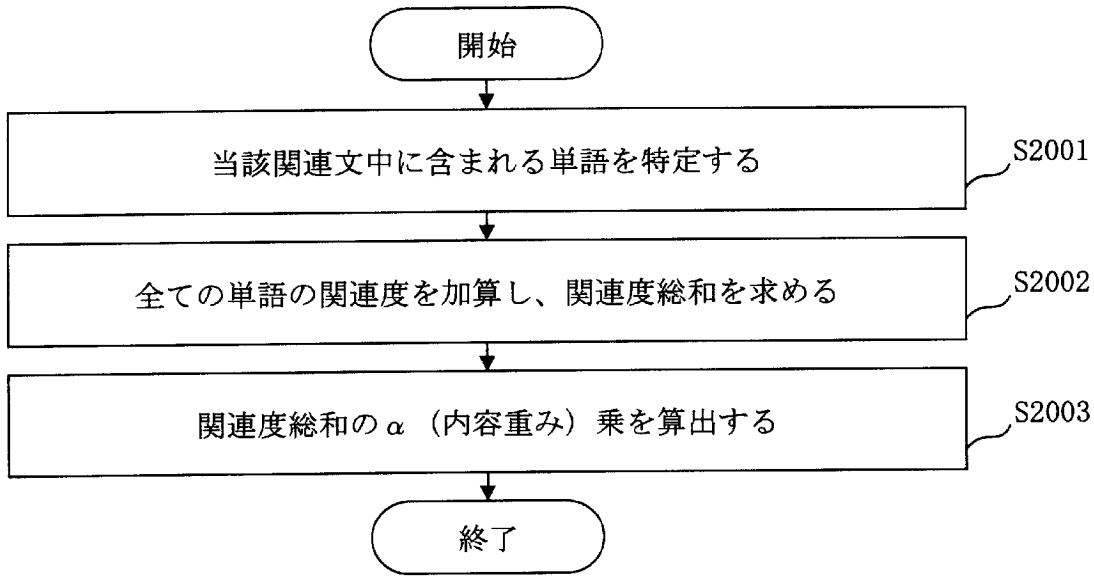
＜関連文書検索処理フロー＞

[図19]



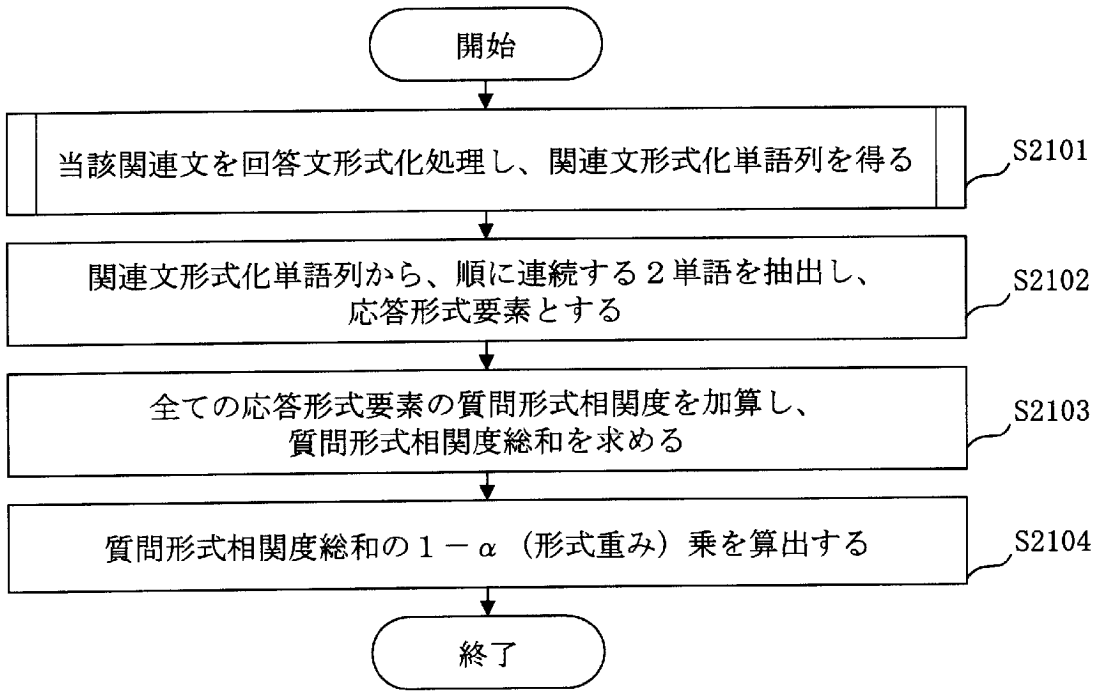
＜文スコア算出処理フロー＞

[図20]



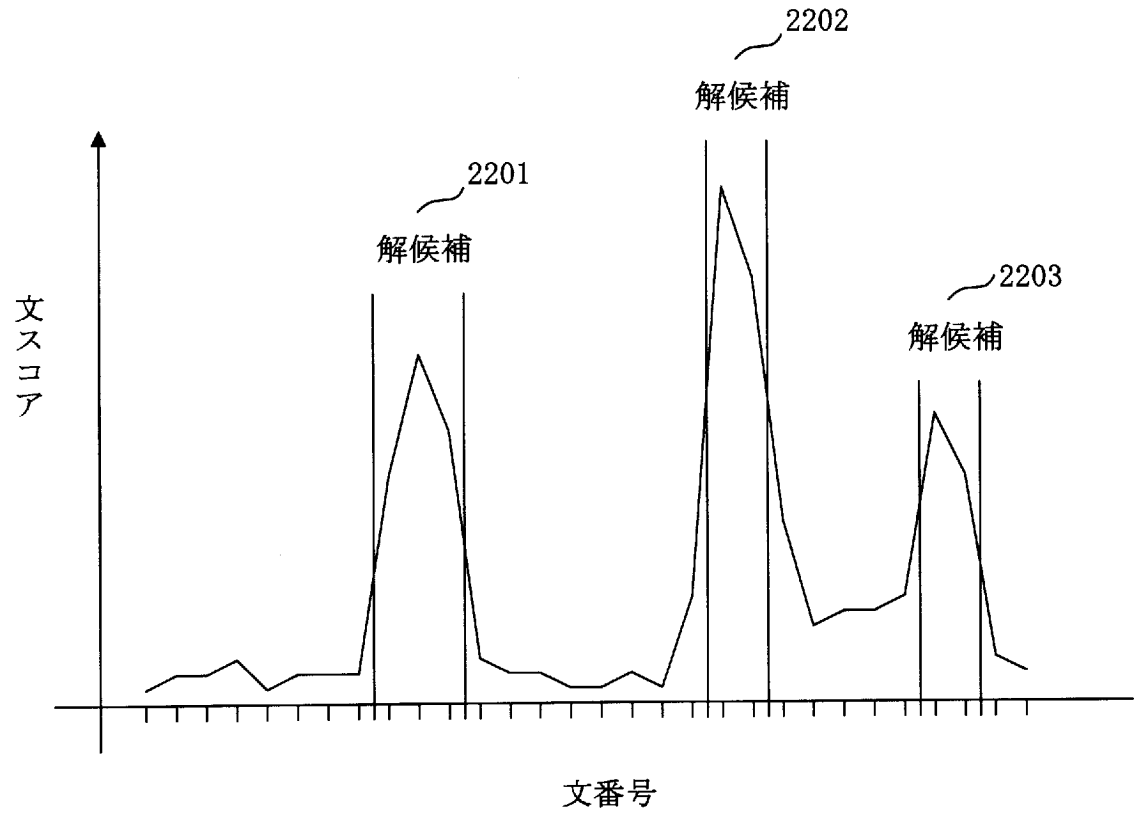
＜内容評価項算出処理フロー＞

[図21]

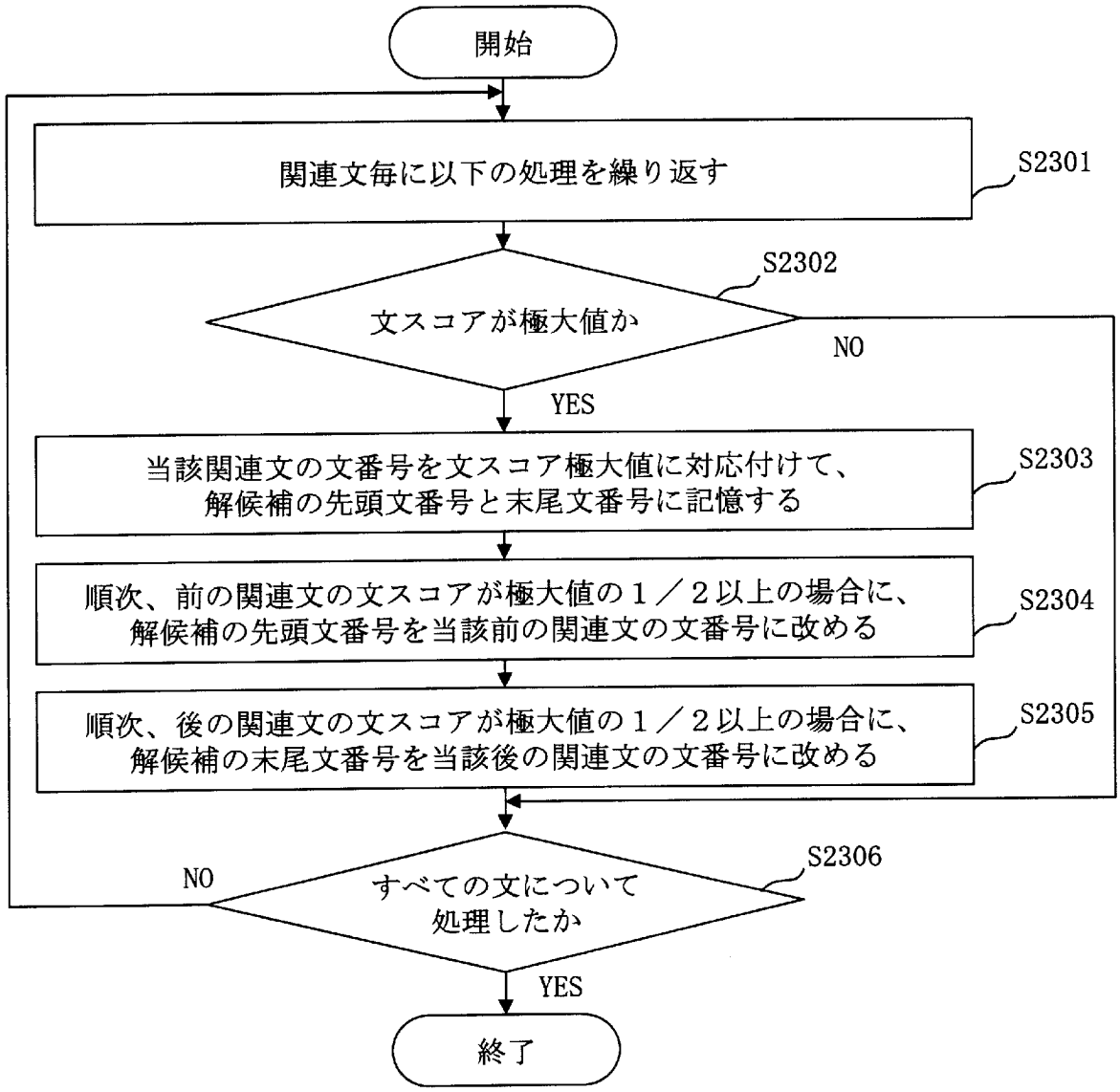


＜形式評価項算出処理フロー＞

[図22]

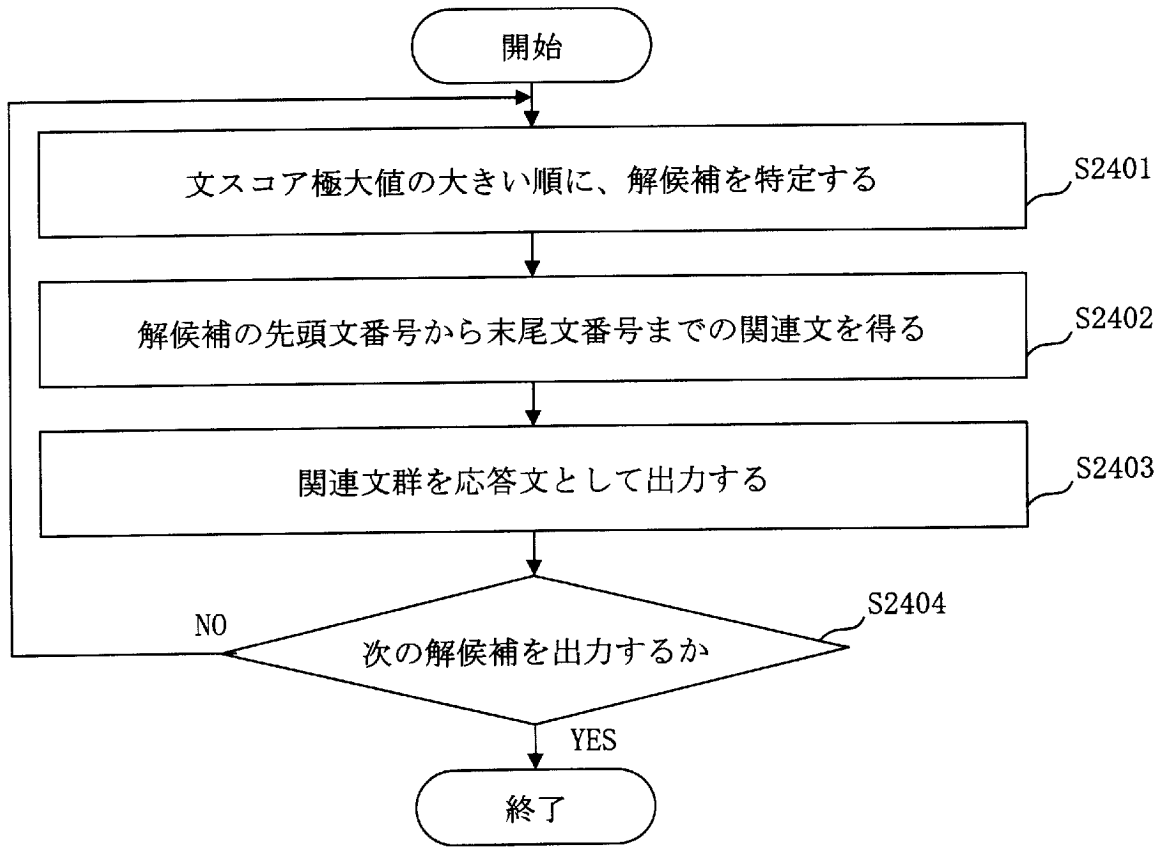


[図23]



＜解候補抽出処理フロー＞

[図24]



＜応答文出力処理フロー＞

[図25]

2501

【正解】 沖縄県に点在する「琉球王国のグスク及び関連遺跡群」、東南アジア、中国、朝鮮半島、日本と経済的政治的交流をもっていたことを示す建造物群であること、失われた琉球王国の遺跡と失われつつある文化的伝統を今に伝える遺産であること、自然崇拝、祖先崇拝という沖縄伝統の信仰形態を今日まで伝えていることなどが評価され、文化遺産に登録されました。

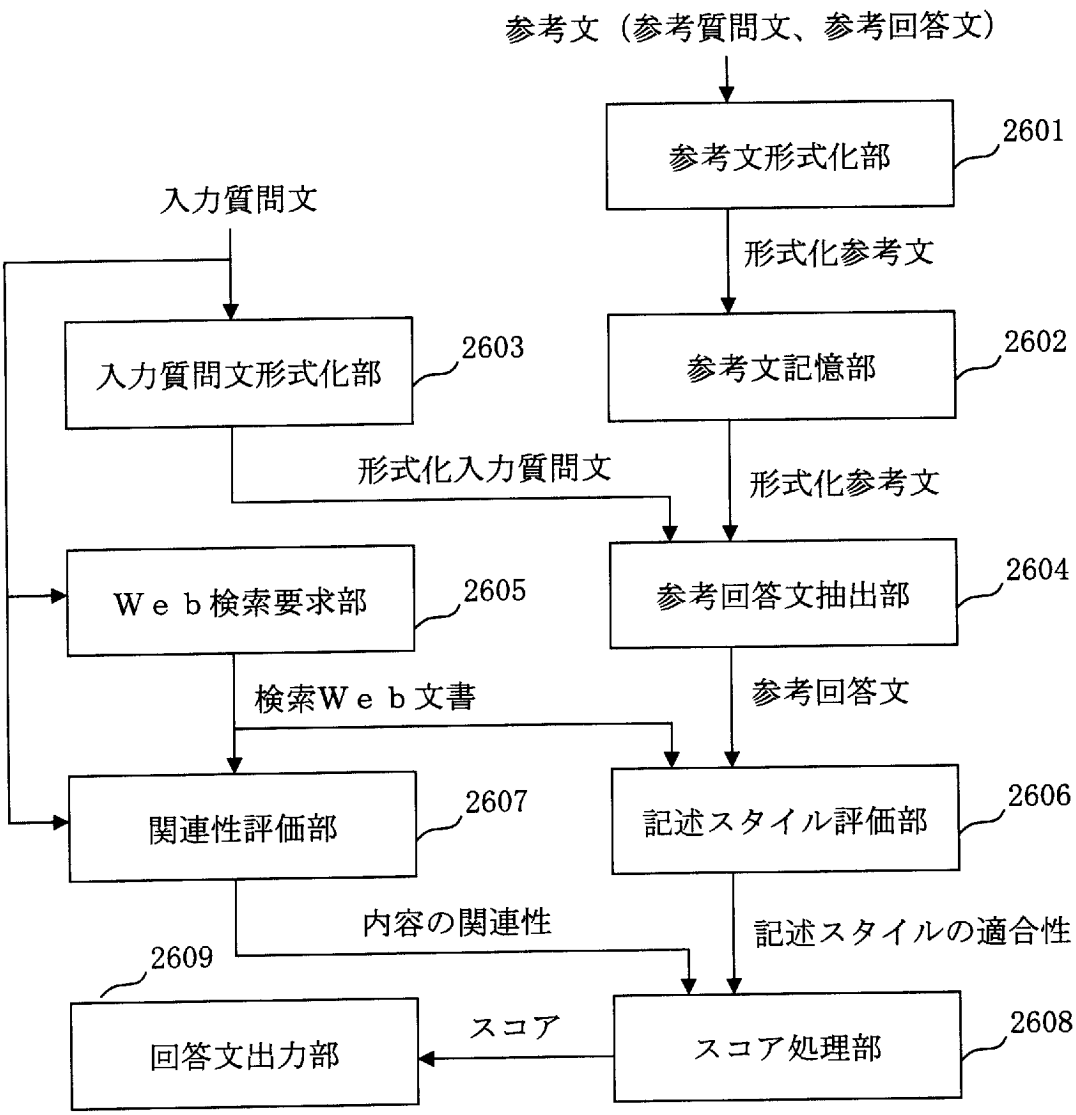
2502

【正解】 選定された理由は・・・一東南アジア、中国、朝鮮半島との長期交流　二祖先崇拝の心が遺産に具現されている 三考古学的価値が高い 四現存の残存遺物と修復された部分が峻別されている これ等は琉球独自の文化と東アジア、東南アジアとの交流が見事にマッチし融合した稀に見る貴重な遺産群である、との評価で選定された。

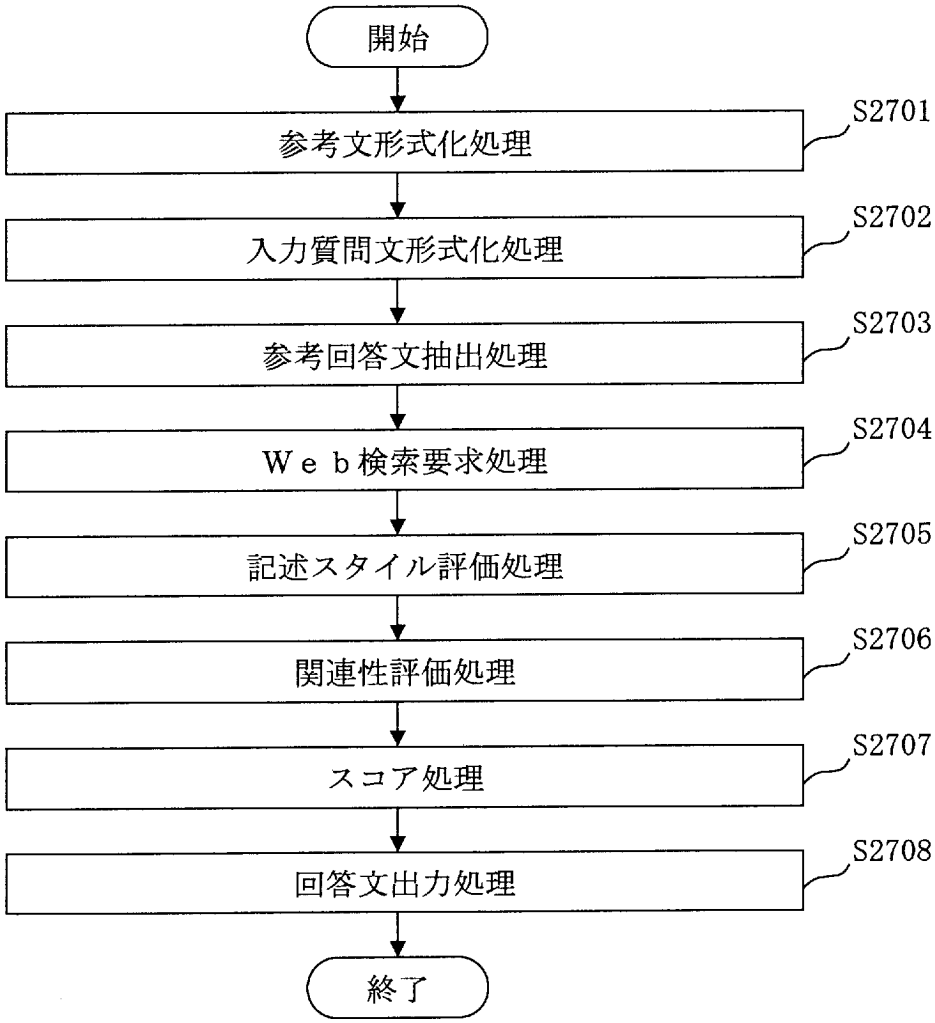
2503

【不正解】２０００年１２月、「琉球王国のグスク及び関連遺産群」として、今帰仁城跡が世界遺産に登録されました。

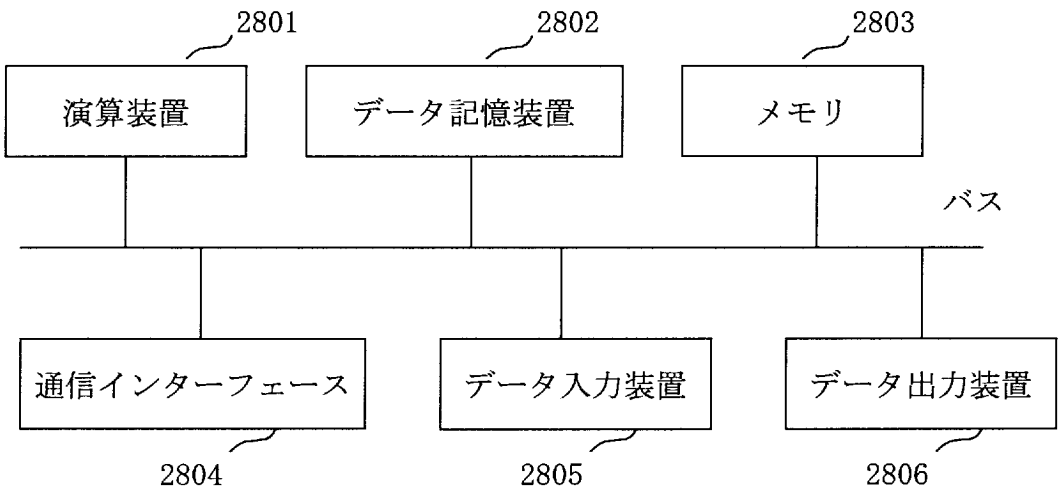
[図26]


参考文（参考質問文、参考回答文）
参考文形式化部
2601
入力質問文
形式化参考文
入力質問文形式化部
2603
参考文記憶部
2602
形式化入力質問文
形式化参考文
Ｗｅｂ検索要求部
2605
参考回答文抽出部
2604
検索Ｗｅｂ文書
参考回答文
関連性評価部
2607
記述スタイル評価部
2606
内容の関連性
記述スタイルの適合性
2609
回答文出力部
スコア
スコア処理部
2608

[図27]


開始
参考文形式化処理
S2701
入力質問文形式化処理
S2702
参考回答文抽出処理
S2703
Ｗｅｂ検索要求処理
S2704
記述スタイル評価処理
S2705
関連性評価処理
S2706
スコア処理
S2707
回答文出力処理
S2708
終了

[図28]


2801
2802
2803
演算装置
データ記憶装置
メモリ
バス
通信インターフェース
データ入力装置
データ出力装置
2804
2805
2806

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2009/054425

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2009
Kokai Jitsuyo Shinan Koho 1971-2009 Toroku Jitsuyo Shinan Koho 1994-2009

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
IEEE Xplore, JSTPlus(JDreamII), NRI Cyber Patent

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2006-39881 A (Nippon Telegraph And Telephone Corp.), 09 February, 2006 (09.02.06), Full text; Figs. 1 to 5 (Family: none)	1-13
A	JP 2004-355550 A (Nippon Telegraph And Telephone Corp.), 16 December, 2004 (16.12.04), Full text; Figs. 1 to 4 (Family: none)	1-13
A	JP 2002-132811 A (Nippon Telegraph And Telephone Corp.), 10 May, 2002 (10.05.02), Full text; Figs. 1, 2 (Family: none)	1-13

[x] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 March, 2009 (27.03.09)	07 April, 2009 (07.04.09)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2009/054425

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Akira KANAI et al., "Fukusu no Web Kensaku Engine o Mochiita Factoid-gata Shitsumon Oto", Information Processing Society of Japan Kenkyu Hokoku (2007-NL-182), 19 November, 2007 (19.11.07), Vol.2007, No.113, pages 101 to 108	1-13
A	Kokoro MOROOKA et al., "Hi Factoid-gata Shitsumon ni Taio shita Sitsumon Oto System", The Associatioin for Natural Language Processing, Dai 13 Kai Nenji Taikai Happyo Ronbunshu [CD-ROM], 2007.03, pages 1 to 4	1-13
A	Hideki Shima, et al., JAVELIN III: Answering Non-Factoid Questions in Japanese, Proc. of NTCIR-6 Workshop Meeting [online], 2007.05.18, p.464-468, [receipt date:2009/3/27], [URL:http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/51.pdf]	1-13
A	Tatsunori MORI, et al., A Monolithic Approach and a Type-by-Type Approach for Non- Factoid Question-answering -Yokohama National University at NTCIR-6 QAC-, Proc. of NTCIR-6 Workshop Meeting [online], 2007.05.18, p.469-476, [receipt date:2009/3/27], [URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/35.pdf]	1-13
A	Masaki Murata, et al., A System for Answering Non-Factoid Japanese Questions by Using Passage Retrieval Weighted Based on Type of Answer, Proc. of NTCIR-6 Workshop Meeting [online], 2007.05.18, p.477-482, [receipt date:2009/3/27], [URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/14.pdf]	1-13
A	Junta Mizuno, et al., Non-factoid Question Answering Experiments at NTCIR-6: Towaards Answer Type Detection for Real World Questions, Proc. of NTCIR-6 Workshop Meeting, 2007.05.18, p.487-492, [receipt date:2009/3/27], [ULR: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/71.pdf]	1-13
P,X	Tatsunori Mori, et al., Answering any class of Japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website, IEEE/WIC/ACM Intn'l Conf. on Web Intelligence and Intelligent Agent Technology[online], 2008.12.12, p.59-65, [receipt date:2009/3/27], [URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4740426]	1-13

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００９／０５４４２５

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００９年
日本国実用新案登録公報	１９９６－２００９年
日本国登録実用新案公報	１９９４－２００９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

IEEE Xplore, JSTPlus(JDreamII), NRI サイバーパテント

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	JP 2006-39881 A (日本電信電話株式会社) 2006.02.09, 全文, 第1-5図 (ファミリーなし)	1-13
A	JP 2004-355550 A (日本電信電話株式会社) 2004.12.16, 全文, 第1-4図 (ファミリーなし)	1-13
A	JP 2002-132811 A (日本電信電話株式会社) 2002.05.10, 全文, 第1,2図 (ファミリーなし)	1-13

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日 ２７．０３．２００９	国際調査報告の発送日 ０７．０４．２００９
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｍ　３２５２ 鈴木　和樹 電話番号　０３－３５８１－１１０１　内線　３５９９

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	金井明、外３名，複数のＷｅｂ検索エンジンを用いたｆａｃｔｏｉｄ型質問応答，情報処理学会研究報告(2007-NL-182)，2007.11.19，第2007巻，第113号，p.101-108	1-13
A	諸岡心、外１名，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[CD-ROM]，2007.03，p.1-4	1-13
A	Hideki Shima、外１名，JAVELIN III: Answering Non-Factoid Questions in Japanese，Proc. of NTCIR-6 Workshop Meeting [online]，2007.05.18，p.464-468，[検索日:2009/3/27]，[ＵＲＬ：http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/51.pdf]	1-13
A	Tatsunori MORI、外５名，A Monolithic Approach and a Type-by-Type Approach for Non-Factoid Question-answering -Yokohama National University at NTCIR-6 QAC-，Proc. of NTCIR-6 Workshop Meeting [online]，2007.05.18，p.469-476，[検索日:2009/3/27]，[ＵＲＬ：http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/35.pdf]	1-13
A	Masaki Murata、外４名，A System for Answering Non-Factoid Japanese Questions by Using Passage Retrieval Weighted Based on Type of Answer，Proc. of NTCIR-6 Workshop Meeting[online]，2007.05.18，p.477-482，[検索日:2009/3/27]，[ＵＲＬ：http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/14.pdf]	1-13
A	Junta Mizuno、外１名，Non-factoid Question Answering Experiments at NTCIR-6: Towaards Answer Type Detection for Real World Questions，Proc. of NTCIR-6 Workshop Meeting，2007.05.18，p.487-492，[検索日:2009/3/27]，[ＵＲＬ：http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/71.pdf]	1-13
P,X	Tatsunori Mori、外２名，Answering any class of Japanese non-factoid question by using the Web and example Q&A pairs from a social Q&A website，IEEE/WIC/ACM Intn'l Conf. on Web Intelligence and Intelligent Agent Technology[online]，2008.12.12，p.59-65，[検索日:2009/3/27]，[ＵＲＬ：http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4740426]	1-13