

明 細 書

発明の名称：類義語判定システム、および類義語判定方法

技術分野

[0001] 本発明は、類義語判定システム、および類義語判定方法に関する。

背景技術

[0002] 本出願は、2021年5月12日に出願された日本特許出願2021-080731号に基づく優先権を主張し、その開示全体を援用して本出願に取り込むものである。

[0003] 特許文献1には、単語の数値表現を計算するためのシステムについて記載されている。上記システムは、単語のシーケンスを含む学習データを用いて複数の分類器および埋め込み関数を学習する。分類器は、入力単語の数値表現を処理することにより単語スコアを生成する。埋め込み関数は、入力単語を受け取り、埋め込み関数パラメータに従って入力単語を高次元空間の数値表現にマッピングする。上記システムは、埋め込み関数の層を使用して単語の語彙内の各単語を処理し、高次元空間内の語彙内の各単語のそれぞれの数値表現を取得し、語彙の各単語を高次元空間の単語に関連付ける。

[0004] 特許文献2には、表記ゆれ候補を精度よく検出することを目的として構成された表記ゆれ検出装置について記載されている。表記ゆれ検出装置は、文書データから用語を抽出し、抽出された用語の任意のペアの類似度を算出し、算出した類似度に基づき、用語のペアが表記ゆれ候補であるかを判定し、表記ゆれ候補である用語のペアに共通に含まれる文字列に基づき表記ゆれ候補をグループ分けする。

先行技術文献

特許文献

[0005] 特許文献1：米国特許第9037464号明細書

特許文献2：特開2012-256197公報

発明の概要

発明が解決しようとする課題

- [0006] 例えば、企業等における製品保守部門においては、製品の故障原因を推定するため、予め蓄積された文書データの中から対応する故障に関連する文書（以下、「製品保守文書」と称する。）を検索することがある。こうした検索においては、ユーザが指定した検索語のみならず当該検索語の類義語についても同時に検索できるようにすることで、検索効率並びに検索精度の向上を図ることができる。
- [0007] ところで、上記のような類義語による同時検索を行うためには、検索対象となる文書データから予め類義語を抽出しておく必要がある。しかし、大量の文書データからの類義語の抽出を人手で行うには大きな工数がかかるため、文書データから類義語を効率よく抽出するための仕組みが必要となる。
- [0008] ここで特許文献1では、単語のシーケンスを含む学習データにより学習した分類器および埋め込み関数を用いて入力単語を高次元空間の数値表現にマッピングし、高次元空間内の語彙内の各単語のそれぞれの数値表現を取得し、語彙の各単語を高次元空間の単語に関連付ける。しかし同技術により高精度な類義語判定を行おうとすれば膨大な量の学習データを用意する必要があり、例えば、製品保守文書のような特定の技術に特化した文書データから類義語を抽出する場合は学習データを十分に確保できず抽出精度の向上が難しい。
- [0009] 特許文献2では、文書データから抽出した用語の任意のペアの類似度を算出し、算出した類似度に基づき、表記ゆれ候補である用語のペアに共通に含まれる文字列に基づき表記ゆれ候補をグループ分けする。しかし、同文献に記載の技術では、文書の種類ごとに人手でルールを調整する必要がある。例えば、対象となる文書が製品保守文書である場合、対象とする製品ごとに文書中で用いられる単語が異なるため、対象とする製品ごとにルールを設定する必要がある、ルールの設定に要する人的負荷が大きい。
- [0010] 本発明はこうした背景に鑑みなされたものであり、文書データから類義語を効率よく高い精度で抽出することが可能な、類義語判定システム、および

類義語判定方法を提供することを目的とする。

課題を解決するための手段

[0011] 上記目的を達成するための本発明の1つは、類義語判定システムであって、プロセッサおよびメモリを有する情報処理装置を用いて構成され、文書データから抽出される複数の単語から選択される2つの単語の組合せである複数の類義語候補の一部について、構成する2つの単語が類義語の関係となるか否かを示す情報である正誤情報を取得し、前記文書データから取得される前記類義語候補についての特徴量と前記正誤情報とに基づき、前記類義語候補を構成する2つの単語が類義語の関係となるか否かを判定するための情報である類義語抽出ルールを生成し、前記正誤情報を取得していない前記類義語候補に前記類義語抽出ルールを適用することにより、構成する2つの単語が類義語の関係となる前記類義語候補を抽出する。

[0012] その他、本願が開示する課題、およびその解決方法は、発明を実施するための形態の欄、および図面により明らかにされる。

発明の効果

[0013] 本発明によれば、文書データから類義語を効率よく高い精度で抽出することができる。

図面の簡単な説明

[0014] [図1]類義語判定システムの概略的な構成を示す図である。

[図2]類義語判定システムの構成に用いる情報処理装置の一例である。

[図3]類義語判定システムの基本的な動作を説明するシステムフロー図である。

[図4]文書情報テーブルの一例である。

[図5]単語カテゴリリストの一例である。

[図6]単語テーブルの一例である。

[図7]類義語候補テーブルの一例である。

[図8]不一致部分列テーブルの一例である。

[図9]閾値テーブルの一例である。

[図10]部分列正誤テーブルの一例である。

[図11]類義語辞書の一例である。

[図12]非類義語辞書の一例である。

[図13]類義語判定処理を説明するフローチャートである。

[図14]単語抽出処理を説明するフローチャートである。

[図15]類義語候補生成処理を説明するフローチャートである。

[図16]類義語抽出ルール適用処理を説明するフローチャートである。

[図17]不一致部分列特定処理を説明するフローチャートである。

[図18]類義語候補正誤判定処理を説明するフローチャートである。

[図19]正誤判定入力画面の一例である。

[図20]類義語抽出ルール生成処理を説明するフローチャートである。

[図21]閾値決定処理を説明するフローチャートである。

[図22]特徴量正誤数分布の一例である。

[図23]部分列正誤テーブル生成処理を説明するフローチャートである。

[図24]再提示要否確認ダイアログの一例である。

発明を実施するための形態

[0015] 以下、発明の実施形態について図面を参照しつつ説明する。以下の記載および図面は、本発明を説明するための例示であって、説明の明確化のため、適宜、省略および簡略化がなされている。本発明は、他の種々の形態でも実施することが可能である。とくに限定しない限り、各構成要素は単数でも複数でも構わない。

[0016] 以下の説明において、同一または類似の構成について同一の符号を付して重複した説明を省略することがある。また、以下の説明において、符号の前に付した「S」の文字は処理ステップを意味する。また、以下の説明では、「テーブル」、「情報」等の表現にて各種情報を説明することがあるが、情報はこれら以外のデータ構造で表現されていてもよい。

[0017] 以下の説明において、2つの単語の組合せのことを「単語ペア」と称する。また以下の説明において、所定のトピックについて記述された、1つ以上

の文や1つ以上の文章が纏まったものを文書と称し、以下で説明する各種の処理は原則として文書を単位として行われるものとして説明するが、処理の単位は必ずしも限定されない。

[0018] 図1に本発明の一実施形態として説明する情報処理システム（以下「類義語判定システム1」と称する）の概略的な構成を示している。同図に示すように、類義語判定システム1は、類義語判定装置100、ユーザ装置2、およびデータ管理装置4を含む。これらはいずれも情報処理装置（コンピュータ）を用いて構成されており、通信媒体5を介して互いに双方向通信が可能な状態で接続されている。通信媒体5は、有線または無線による通信機能を提供する通信基盤であり、例えば、LAN（Local Area Network）、WAN（Wide Area Network）、インターネット、専用線、各種の公衆通信網等の各種通信ネットワーク、USB（Universal Serial Bus）規格に準拠した各種通信機器等である。

[0019] 類義語判定装置100は、文書データから抽出した単語ペアにおける2つの単語が類義語または非類義語の関係にあるかを判定し、類義語の関係にあると判定された単語ペアについては類義語辞書に登録し、非類義語の関係にあると判定された単語ペアについては非類義語辞書に登録する。単語ペアの抽出元の文書データは、例えば、OT知識（OT：Operational Technology）が記述されている製品保守文書等を電子化したデータである。類義語判定装置100によって生成された類義語辞書は、例えば、OT知識から有用な情報を効率よく検索してユーザに提供するサービスに用いられる。

[0020] 同図に示すように、類義語判定装置100は、主な機能として、記憶部110、単語抽出部130、類義語候補生成部140、類義語抽出ルール適用部150、類義語候補正誤判定部160、不一致部分列特定部170、および類義語抽出ルール生成部180を備える。また、類義語抽出ルール生成部180は、閾値決定部181および部分列正誤テーブル生成部182を含む。

[0021] 記憶部110は、主な情報（データ）として、文書情報テーブル111、

単語カテゴリリスト 112、単語カテゴリ判定モデル 113、単語テーブル 114、類義語候補テーブル 115、不一致部分列テーブル 116、閾値テーブル 117、部分列正誤テーブル 118、類義語辞書 121、および非類義語辞書 122 を記憶する。これらの詳細については後述する。

[0022] ユーザ装置 2 は、類義語判定装置 100 が参照もしくは更新する各種の情報を管理するためのユーザインタフェース（画面（画像）表示装置や音声入出力装置等）を提供する。ユーザ装置 2 は、例えば、類義語候補テーブル 115、類義語辞書 121、および非類義語辞書 122 をユーザが参照もしくは編集するためのユーザインタフェースを提供する。また、ユーザ装置 2 は、ユーザインタフェースを介して、類義語候補となる単語ペアが管理されている類義語候補テーブル 115 の単語ペアが類義語であるか否かを示す情報（以下、「正誤情報」と称する。）をユーザから受け付け、受け付けた正誤情報を、通信媒体 5 を介して類義語判定装置 100 に送信する。

[0023] データ管理装置 4 は、データ管理通信部 41 を備える。データ管理通信部 41 は、単語ペアの抽出元となる文書データを文書情報テーブル 42 に管理する。また、データ管理通信部 41 は、類義語判定装置 100 と通信し、類義語判定装置 100 に文書データを適宜提供（送信）する。データ管理装置 4 は、文書情報テーブル 42 に管理される文書データを、例えば、通信媒体 5 を介して取得する。また、上記の文書データは、ユーザ装置 2 が提供するユーザインタフェースを介してユーザが登録することもできる。

[0024] 図 2 に、類義語判定システム 1（類義語判定装置 100、ユーザ装置 2、データ管理装置 4 等）の構成に用いる情報処理装置の一例を示す。例示する情報処理装置 10 は、プロセッサ 11、主記憶装置 12、補助記憶装置 13、入力装置 14、出力装置 15、および通信装置 16 を備える。情報処理装置 10 の例として、パーソナルコンピュータ、サーバ装置、スマートフォン、タブレット等がある。

[0025] 情報処理装置 10 は、その全部または一部が、例えば、クラウドシステムによって提供される仮想サーバのように、仮想化技術やプロセス空間分離技

術等を用いて提供される仮想的な情報処理資源を用いて実現されるものであってもよい。また、情報処理装置 10 によって提供される機能の全部または一部は、例えば、クラウドシステムが A P I (Application Programming Interface) 等を介して提供するサービスによって実現してもよい。また、情報処理装置 10 によって提供される機能の全部または一部は、例えば、S a a S (Software as a Service)、P a a S (Platform as a Service)、I a a S (Infrastructure as a Service) 等を利用して実現されるものであってもよい。

[0026] 類義語判定装置 100 とユーザ装置 2 は、同じ情報処理装置 10 (共通のハードウェア) によって実現してもよい。また、類義語判定装置 100 は、例えば、通信可能に接続された複数の情報処理装置 10 を用いて実現してもよい。

[0027] 同図に示すプロセッサ 11 は、例えば、C P U (Central Processing Unit)、M P U (Micro Processing Unit)、G P U (Graphics Processing Unit)、F P G A (Field Programmable Gate Array)、A S I C (Application Specific Integrated Circuit)、A I (Artificial Intelligence) チップ等を用いて構成されている。

[0028] 主記憶装置 12 は、プログラムやデータを記憶する装置であり、例えば、R O M (Read Only Memory)、R A M (Random Access Memory)、不揮発性メモリ (N V R A M (Non Volatile RAM)) 等である。

[0029] 補助記憶装置 13 は、例えば、S S D (Solid State Drive)、ハードディスクドライブ、光学式記憶装置 (C D (Compact Disc)、D V D (Digital Versatile Disc) 等)、ストレージシステム、I C カード、S D カードや光学式記録媒体等の記録媒体の読取／書込装置、クラウドサーバの記憶領域等である。補助記憶装置 13 には、記録媒体の読取装置や通信装置 16 を介してプログラムやデータを読み込むことができる。補助記憶装置 13 に格納 (記憶) されているプログラムやデータは主記憶装置 12 に随時読み込まれる。

[0030] 入力装置 14 は、外部からの入力を受け付けるインタフェースであり、例

えば、キーボード、マウス、タッチパネル、カードリーダー、ペン入力方式のタブレット、音声入力装置等である。

- [0031] 出力装置 15 は、処理経過や処理結果等の各種情報を出力するインタフェースである。出力装置 15 は、例えば、上記の各種情報を可視化する表示装置（液晶モニタ、LCD (Liquid Crystal Display)、グラフィックカード等）、上記の各種情報を音声化する装置（音声出力装置（スピーカ等））、上記の各種情報を文字化する装置（印字装置等）である。尚、例えば、情報処理装置 10 が通信装置 16 を介して他の装置との間で情報の入力や出力を行う構成としてもよい。
- [0032] 入力装置 14 と出力装置 15 は、ユーザとの間での対話処理（情報の受け付け、情報の提示等）を実現するユーザインタフェースを構成する。
- [0033] 通信装置 16 は、他の装置との間の通信を実現する装置である。通信装置 16 は、通信媒体 5 を介して他の装置との間の通信を実現する、有線方式または無線方式の通信インタフェースであり、例えば、NIC (Network Interface Card)、無線通信モジュール、USB モジュール等である。
- [0034] 情報処理装置 10 には、例えば、オペレーティングシステム、ファイルシステム、DBMS (DataBase Management System)（リレーショナルデータベース、NoSQL 等）、KVS (Key-Value Store) 等が導入されていてもよい。
- [0035] 類義語判定装置 100、ユーザ装置 2、およびデータ管理装置 4 の機能は、夫々のプロセッサ 11 が、夫々の主記憶装置 12 に格納されているプログラムを読み出して実行することにより、もしくは、これらの装置を構成するハードウェア（FPGA、ASIC、AI チップ等）自体によって実現される。
- [0036] 類義語判定装置 100 が提供する各種の機能は、例えば、テキストデータマイニング等の公知の各種データマイニング手法、公知の各種自然言語処理手法（形態素解析、構文解析、意味解析、文脈解析、特徴量抽出、単語分散表現、固有表現抽出、テキスト分類、系列ラベリング）、公知の各種機械学

習手法（深層学習（DNN（Deep Neural Network）、RNN（Recurrent Neural Network）等）を用いて実現される。類義語判定装置100は、前述した各種の情報（データ）を、例えば、データベースのテーブルやファイルシステムが管理するファイルとして記憶する。

[0037] 図3は、類義語判定システム1の基本的な動作を説明するシステムフロー図である。以下、同図並びに図1とともに、類義語判定システム1の基本的な動作について説明する。類義語判定システム1は、単語ペアを構成する2つの単語が類義語か非類義語かを判定する基準として、ユーザから受け付けた正誤情報や類義語抽出ルール生成部180によって生成される類義語抽出ルール（閾値テーブル117、部分列正誤テーブル118）を用いる。

[0038] 同図に示すように、類義語判定装置100の単語抽出部130は、文書情報テーブル111から単語を抽出し、抽出した単語が属するカテゴリを単語カテゴリ判定モデル113により特定する。単語抽出部130は、例えば、文書情報に固有表現抽出（NER:Named Entity Recognition）技術を適用することにより単語を抽出する。単語抽出部130は、抽出した単語と単語が属するカテゴリ、単語が特定した上記カテゴリに属する確率（以下、「カテゴリ関連確率」と称する。）、および文書情報テーブル中の抽出した単語の抽出元の文章（テキストデータ）を単語テーブル114に登録する。

[0039] 類義語候補生成部140は、単語テーブル114に管理されている、カテゴリが同一である2つの単語の組合せである単語ペア（以下、「類義語候補」とも称する。）について、当該単語ペアを構成する2つの単語の関係性を示す特徴量（以下、「関係性特徴量」と称する。）を求め、単語ペアと当該単語ペアの関係性特徴量とを対応づけて類義語候補テーブル115に格納する。類義語候補生成部140は、関係性特徴量として、例えば、文書情報テーブル111の文書データから機械学習モデル(word2vecなど)を適用することにより取得される単語ペアの共起度数、単語ペアの編集距離、単語ペアのカテゴリ関連確率、単語ペアの出現回数、単語ペアの各単語の抽出元の文章（テキストデータ）等を用いる。尚、関係性特徴量は必ずしもこれらに限定

されない。また、類義語候補テーブル 115 の内容は、ユーザ装置 2 が提供するユーザインタフェースを介してユーザが設定してもよい。

[0040] 類義語抽出ルール適用部 150 は、類義語候補テーブル 115 の単語ペアを構成する 2 つの単語が、類義語または非類義語の関係にあるかを判定する。類義語抽出ルール生成部 180 は、上記判定を類義語抽出ルール（閾値テーブル 117、部分列正誤テーブル 118）を用いて行う。具体的には、類義語抽出ルール適用部 150 は、まず類義語候補テーブル 115 の単語ペアについて、その単語間で一致していない部分の文字列（以下、「不一致部分列」と称する。）の組合せ（以下、「不一致部分列ペア」と称する。）を特定する。尚、上記の特定は、具体的には、類義語抽出ルール適用部 150 が不一致部分列特定部 170 に指示することにより行われる。続いて、類義語抽出ルール適用部 150 は、部分列正誤テーブル 118 を参照し、特定した上記の不一致部分列ペアの正誤（類義語の関係か非類義語の関係か）を確認し、正の判定（類義語である旨の判定）を受けていた場合は単語ペアを類義語辞書 121 に登録し、誤の判定（非類義語である旨の判定）を受けていた場合は非類義語辞書 122 に単語ペアを登録する。また、類義語抽出ルール適用部 150 は、類義語候補テーブル 115 の単語ペアについて、当該単語ペアの関係性特徴量と夫々に対応する閾値テーブル 117 の閾値とを比較し、閾値テーブルの値を下回る関係性特徴量がある場合は、当該単語ペアを非類義語辞書 122 に登録する。尚、本例ではこのように 1 つでも閾値テーブルの値を下回る関係性特徴量がある場合に単語ペアを非類義語辞書 122 に登録するとするが、非類義語とするか否かの判定の条件は必ずしも限定されない。

[0041] 類義語候補正誤判定部 160 は、類義語候補テーブル 115 の単語ペアを構成する 2 つの単語が類義語の関係にあるか非類義語の関係にあるかを示す情報（以下、「正誤情報」と称する。）を取得し、類義語の関係にある場合は当該単語ペアを類義語辞書 121 に登録し、非類義語の関係にある場合は当該単語ペアを非類義語辞書 122 に登録する。本実施形態では、ユーザ装

置 2 に類義語候補テーブル 115 を提示しつつ単語ペアの正誤情報をユーザから受け付ける。尚、正誤情報の取得方法は必ずしも限定されない。例えば、他の情報処理システムにより生成された正誤情報を利用してもよい。類義語候補正誤判定部 160 は、単語ペアの正誤情報に基づき、単語ペアに対応する類義語候補テーブル 115 に管理されている正誤情報を更新する。

[0042] 不一致部分列特定部 170 は、不一致部分列ペアを特定し、特定した不一致部分列ペアを不一致部分列テーブル 116 に登録する。例えば、単語ペアの各単語が「SIPノズル」と「真空ノズル」である場合、不一致部分列特定部 170 は、「SIP」と「真空」を不一致部分列ペアとして特定し、特定した不一致部分列ペアを不一致部分列テーブル 1167 に登録する。

[0043] 類義語抽出ルール生成部 180 は、類義語候補テーブル 115 に管理されている各単語ペアの関係性特微量と正誤情報とに基づき、類義語抽出ルール（閾値テーブル 117、部分列正誤テーブル 118）を生成する。具体的には、類義語抽出ルール生成部 180 の閾値決定部 181 は、カテゴリごとに、関係性特微量の値と正誤の数の関係（以下、「特微量正誤数分布」と称する。）を求め、特微量正誤数分布に基づき、関係性特微量について設定する閾値を決定する。閾値決定部 181 は、決定した閾値を閾値テーブル 117 に登録する。

[0044] 部分列正誤テーブル生成部 182 は、類義語候補テーブル 115 の単語ペアについて、不一致部分列特定部 170 に指示して不一致部分列ペアを特定し、特定した不一致部分列ペアと当該単語ペアの正誤情報とを対応づけたレコードを部分列正誤テーブル 118 に登録する。

[0045] 続いて、記憶部 110 が管理する主な情報（データ）について具体的に説明する。

[0046] 図 4 は、図 1 に示した記憶部 110 が管理する文書情報テーブル 111 の一例である。文書情報テーブル 111 には、単語ペアの抽出元である文書データ（例えば、テキストデータや文書の内容を所定のデータフォーマットで記述したデータ）が管理される。例示する文書情報テーブル 111 は、文書

ＩＤ１１１１およびテキスト１１１２の各項目を有する一つ以上のレコードで構成される。

[0047] 上記項目のうち、文書ＩＤ１１１１には、文書データの識別子（以下、「文書ＩＤ」と称する。）が格納される。テキスト１１１２には、文書データの実体が格納される。尚、テキスト１１１２に文書データの所在のみを格納し、文書データの実体については上記所在で特定される記憶領域（例えば、類義語判定装置１００と通信可能に接続されたストレージ装置等）に管理するようにしてもよい。

[0048] 図５は、記憶部１１０が管理する単語カテゴリリスト１１２の一例である。単語カテゴリリスト１１２には、単語の分類先となるカテゴリの一覧（リスト）が管理される。例示する単語カテゴリリスト１１２は、カテゴリの一覧が格納される項目であるカテゴリ１１２１を有する。単語カテゴリリスト１１２の内容は、例えば、ユーザ装置２が提供するユーザインタフェースを介してユーザが設定する。

[0049] 図１に示した記憶部１１０が管理する単語カテゴリ判定モデル１１３は、文書情報テーブル１１１から抽出しようとする単語が単語カテゴリリスト１１２のいずれのカテゴリに属するかを判定するモデルである。単語カテゴリ判定モデル１１３は、入力された単語が単語カテゴリリスト１１２の各カテゴリに属する確率（以下、「カテゴリ関連確率」と称する。）を求め、求めたカテゴリ関連確率が最も大きいカテゴリを単語が属するカテゴリと判定し、判定した単語と当該単語のカテゴリ関連確率を出力する。単語カテゴリ判定モデル１１３は、例えば、行列式や数式で表される。単語カテゴリ判定モデル１１３は、例えば、公知の各種自然言語処理手法（形態素解析、構文解析、意味解析、文脈解析、特徴量抽出、単語分散表現、固有表現抽出、テキスト分類、系列ラベリング）や、公知の各種機械学習手法（深層学習（ＤＮＮ（Deep Neural Network）、ＲＮＮ（Recurrent Neural Network）等）を用いて実現される。但し、単語カテゴリ判定モデル１１３の実現方法は必ずしも限定されない。

- [0050] 図6は、図1に示した記憶部110が管理する単語テーブル114の一例である。同図に示すように、単語テーブル114には、単語1141、単語カテゴリ1142、カテゴリ関連確率1143、出現回数1144、および抽出元テキスト1145の各項目を有する複数のレコードで構成される。単語テーブル114の一つのレコードは一つの単語に対応する。
- [0051] 上記項目のうち、単語1141には、文書情報テーブル111のテキスト1112から単語抽出部130が抽出した単語のテキストデータが格納される。単語カテゴリ1142には、単語カテゴリ判定モデル113により判定された、当該単語が属するカテゴリが格納される。カテゴリ関連確率1143には、単語カテゴリ判定モデル113が求めた当該単語のカテゴリ関連確率が格納される。出現回数1144には、当該単語の抽出元の文書データにおける出現回数が格納される。抽出元テキスト1145には、当該単語の抽出元の文書データであるテキストデータが格納される。
- [0052] 図7は、図1に示した記憶部110が管理する類義語候補テーブル115の一例である。類義語候補テーブル115には、類義語候補である単語ペアの情報が管理される。単語ペアは、単語テーブル114の単語から抽出された同一のカテゴリに属する2つの単語の組合せである。同図に示すように、例示する類義語候補テーブル115は、単語A1151、単語B1152、正誤情報1153、単語カテゴリ1154、共起度数1155、編集距離1156、カテゴリ関連確率1157、出現回数1158、および抽出元テキスト1159の各項目を有する複数のレコードで構成される。
- [0053] 上記項目のうち、単語A1151および単語B1152には、夫々、類義語候補となる単語ペアの要素（以下、「単語A」、「単語B」と称する。）が格納される。単語カテゴリ1154には、単語テーブル114から取得される、単語Aおよび単語Bが属するカテゴリが格納される。尚、本例では、単語Bが単語Aの類義語の候補であるものとする。
- [0054] 正誤情報1153には、当該単語ペアを構成する2つの単語（単語Aと単語B）が類義語の関係にあるか否かを示す情報（正誤情報）が格納される。

正誤情報 1 1 5 3 の内容は、ユーザ装置 2 が提供するユーザインタフェースを介してユーザから受け付ける。正誤情報が未取得である場合は正誤情報 1 1 5 3 にその旨を示す情報（例えば「不明」）が格納される。

[0055] 共起度数 1 1 5 5、編集距離 1 1 5 6、カテゴリ関連確率 1 1 5 7、出現回数 1 1 5 8、および抽出元テキスト 1 1 5 9 には、関係性特徴量（共起度数、編集距離、カテゴリ関連確率、出現回数、抽出元テキスト）の具体的な値が格納される。共起度数 1 1 5 5 には、機械学習モデル等を用いて算出された単語 A と単語 B の共起度が格納される。編集距離 1 1 5 6 には、当該単語ペアの編集距離を単語 A と単語 B の長さの合計を用いて正規化した値が格納される。カテゴリ関連確率 1 1 5 7 には、単語テーブル 1 1 4 から取得される、単語 A および単語 B の夫々のカテゴリ関連確率が格納される。出現回数 1 1 5 8 には、単語テーブル 1 1 4 から取得される、単語 A および単語 B の夫々の出現回数が格納される。抽出元テキスト 1 1 5 9 には、単語テーブル 1 1 4 から取得される、単語 A および単語 B の夫々の抽出元の文書データ（テキストデータ）が格納される。

[0056] 図 8 は、図 1 に示した記憶部 1 1 0 が管理する不一致部分列テーブル 1 1 6 の一例である。不一致部分列テーブル 1 1 6 には、不一致部分列ペア（単語ペアを構成する単語 A と単語 B の夫々の不一致部分列のテキストデータ）が格納される。同図に示すように、例示する不一致部分列テーブル 1 1 6 は、部分列 A 1 1 6 1、部分列 B 1 1 6 2 の各項目を有する複数のレコードで構成される。不一致部分列テーブル 1 1 6 の一つのレコードは一つの単語ペアに対応する。

[0057] 上記項目のうち、部分列 A 1 1 6 1 および部分列 B 1 1 6 2 には、不一致部分列特定部 1 7 0 により特定された、当該単語ペアを構成する 2 つの単語の間で一致する部分列（以下、「一致部分列」と称する。）を削除することにより残った夫々の単語の部分列（不一致部分列）が格納される。同図の例では、例えば、単語 A 「S I P ノズル」と単語 B 「真空ノズル」からなる単語ペアについて、一致部分列「ノズル」を各単語から削除することにより残

った不一致部分列「S I P」と「真空」が、夫々、部分列A 1 1 6 1、部分列B 1 1 6 2に格納されている。

[0058] 図9は、図1に示した記憶部110が管理する閾値テーブル117の一例である。閾値テーブル117には、閾値決定部181により決定された、カテゴリごとの各関係性特徴量の閾値が格納される。同図に示すように、例示する閾値テーブル117は、単語カテゴリ1171、カテゴリ関連確率閾値1172、出現回数閾値1173、共起度数閾値1174、および編集距離閾値1175の各項目を有する複数のレコードで構成される。閾値テーブル117の一つのレコードは一つのカテゴリに対応する。

[0059] 上記項目のうち、単語カテゴリ1171には、単語カテゴリリスト112におけるカテゴリの一つが格納される。カテゴリ関連確率閾値1172、出現回数閾値1173、共起度数閾値1174、および編集距離閾値1175には、夫々、当該カテゴリについて設定された関係性特徴量（カテゴリ関連確率閾値、単語の出現回数の閾値、共起度数の閾値、単語間の編集距離の閾値）の閾値が格納される。

[0060] 図10は、図1に示した記憶部110が管理する部分列正誤テーブル118の一例である。部分列正誤テーブル118には、不一致部分列テーブル116の不一致部分列ペアの各部分列（部分列A 1 1 6 1と部分列B 1 1 6 2）について、両者に類似性があるか否かを判定した結果が格納される。同図に示すように、部分列正誤テーブル118は、単語カテゴリ1181、部分列A 1 1 8 2、部分列B 1 1 8 3、および正誤情報1184の各項目を有する複数のレコードで構成される。部分列正誤テーブル118の一つのレコードは不一致部分列ペアの一つに対応する。

[0061] 上記項目のうち、単語カテゴリ1181には、当該不一致部分列ペアを構成する各部分列の抽出元の単語が属するカテゴリが格納される。部分列A 1 1 8 2および部分列B 1 1 8 3には、部分列の組合せを構成する各部分列が格納される。正誤情報1184には、不一致部分列ペアが類似性を有するか否かを判定した結果（正誤情報）が格納される。

- [0062] 図 1 1 は、図 1 に示した記憶部 1 1 0 が管理する類義語辞書 1 2 1 の一例である。同図に示すように、類義語辞書 1 2 1 は、単語 1 2 1 1、類義語 1 2 1 2、および単語カテゴリ 1 2 1 3 の各項目を有する複数のレコードで構成される。類義語辞書 1 2 1 の一つのレコードは、一つの単語ペア（ある単語と当該単語の類義語との組合せ）に対応する。
- [0063] 上記項目のうち、単語 1 2 1 1 には、単語テーブル 1 1 4 の単語の一つが格納される。類義語 1 2 1 2 には、当該単語の類義語が格納される。単語カテゴリ 1 2 1 3 には、当該単語および当該類義語が属するカテゴリが格納される。
- [0064] 図 1 2 は、図 1 に示した記憶部 1 1 0 が管理する非類義語辞書 1 2 2 の一例である。同図に示すように、非類義語辞書 1 2 2 は、単語 1 2 2 1、非類義語 1 2 2 2、および単語カテゴリ 1 2 2 3 の各項目を有する複数のレコードで構成される。非類義語辞書 1 2 2 の一つのレコードは、一つの単語ペア（ある単語と当該単語の非類義語との組合せ）に対応する。
- [0065] 上記項目のうち、単語 1 2 2 1 には、単語テーブル 1 1 4 におけるある単語が格納される。非類義語 1 2 2 2 には、当該単語の非類義語が格納される。単語カテゴリ 1 2 2 3 には、当該単語と当該非類義語が属するカテゴリが格納される。
- [0066] 尚、ユーザは、類義語判定システム 1 が提供するユーザインタフェースを介して、類義語辞書 1 2 1 や非類義語辞書 1 2 2 の内容の参照、編集等を行うことができる。
- [0067] 続いて、類義語判定システム 1 において行われる処理について説明する。
- [0068] 図 1 3 は、類義語判定システム 1 が文書情報テーブル 1 1 1 に管理されている文書データに基づき類義語辞書 1 2 1 と非類義語辞書 1 2 2 を生成する処理（以下、「類義語判定処理 S 1 3 0 0」と称する。）を説明するフローチャートである。以下、同図とともに類義語判定処理 S 1 3 0 0 について説明する。類義語判定処理 S 1 3 0 0 は、例えば、ユーザ装置 2 を介してユーザから明示的な指示があったことや予めスケジューリングされたタイミング

が到来したこと等を契機として開始される。

[0069] 同図に示すように、まず単語抽出部 130 が、文書情報テーブル 111 に管理されている文書データから単語を抽出し、抽出した単語を単語テーブル 114 に登録する処理（以下、「単語抽出処理 S1311」と称する。）を行う。単語抽出処理 S1311 の詳細については後述する。

[0070] 続いて、類義語候補生成部 140 が、単語テーブル 114 の同じカテゴリに属する 2 つの単語の組合せ（単語ペア）について特徴量を求め、単語ペアと求めた特徴量とを類義語候補テーブル 115 に登録する処理（以下、「類義語候補生成処理 S1312」と称する。）を行う。類義語候補生成処理 S1312 の詳細については後述する。

[0071] 続いて、類義語抽出ルール適用部 150 が、類義語候補テーブル 115 の単語ペアを構成する 2 つの単語について、類義語抽出ルール（閾値テーブル 117、部分列正誤テーブル 118）を用いて類義語または非類義語の関係にあるかを判定し、判定の結果に応じて単語ペアを類義語辞書 121 または非類義語辞書 122 に登録する処理（以下、「類義語抽出ルール適用処理 S1313」と称する。）を行う。類義語抽出ルール適用処理 S1313 の詳細については後述する。

[0072] 続いて、類義語候補正誤判定部 160 が、類義語候補テーブル 115 の単語ペアについてユーザから正誤情報（類義語か非類義語かを示す情報）を取得する処理（以下、「類義語候補正誤判定処理 S1314」と称する。）を行う。類義語候補正誤判定処理 S1314 の詳細については後述する。

[0073] 続いて、類義語抽出ルール生成部 180 が、類義語候補テーブル 115 の単語ペアの正誤情報に基づき、類義語抽出ルール（閾値テーブル 117、部分列正誤テーブル 118）を生成する処理（以下、「類義語抽出ルール生成処理 S1315」と称する。）を行う。類義語抽出ルール生成処理 S1315 の詳細については後述する。

[0074] 続く S1316 の処理については後述する。尚、以上の処理のうち、S1311～S1312 の処理については、例えば、S1313～S1316 と

は独立したタイミングで実行するようにしてもよい。例えば、S 1 3 1 1 ～ S 1 3 1 2 の処理は、文書情報テーブル 1 1 1 が更新されたことを契機として実行し、一方、S 1 3 1 3 ～ S 1 3 1 6 の処理については、例えば、ユーザ装置 2 を介してユーザから類義語の抽出要求（類義語辞書 1 2 1 の作成要求）を受け付けたことを契機として実行するようにしてもよい。

[0075] 図 1 4 は、図 1 3 に示した単語抽出処理 S 1 3 1 1 の詳細を説明するフローチャートである。以下、同図とともに単語抽出処理 S 1 3 1 1 について説明する。

[0076] まず、単語抽出部 1 3 0 が、文書情報テーブル 1 1 1 を取得する（S 1 4 1 1）。

[0077] 続いて、単語抽出部 1 3 0 は、文書情報テーブル 1 1 1 のレコードを一つ選択する（S 1 4 1 2）。

[0078] 続いて、単語抽出部 1 3 0 は、選択したレコードのテキスト 1 1 1 2 に格納されているテキストデータから単語を抽出する。尚、単語抽出部 1 3 0 は、例えば、上記テキストデータについて形態素解析を行うことにより単語を抽出する。そして単語抽出部 1 3 0 は、抽出した単語の一つ（以下、「単語 W」と称する。）を選択する（S 1 4 1 3）。

[0079] 続いて、単語抽出部 1 3 0 は、単語カテゴリ判定モデル 1 1 3 を取得する（S 1 4 1 4）。

[0080] 続いて、単語抽出部 1 3 0 は、単語カテゴリリスト 1 1 2 を取得する（S 1 4 1 5）。

[0081] 続いて、単語抽出部 1 3 0 は、単語カテゴリ判定モデル 1 1 3 と単語カテゴリリスト 1 1 2 を用いて、単語 W が属するカテゴリと当該カテゴリについての単語 W のカテゴリ関連確率を算出する（S 1 4 1 6）。

[0082] 続いて、単語抽出部 1 3 0 は、単語 W が文書情報テーブル 1 1 1 の選択中のレコードのテキスト 1 1 1 2 に格納されているテキストデータにおける出現回数を求める（S 1 4 1 7）。

[0083] 続いて、単語抽出部 1 3 0 は、単語 W、S 1 4 1 6 で求めたカテゴリとカ

テゴリ関連確率、S 1 4 1 7で求めた出現回数、および単語Wの抽出元の文書情報テーブル1 1 1のテキスト1 1 1 2に格納されているテキストデータに対応する各項目（単語1 1 4 1、単語カテゴリ1 1 4 2、カテゴリ関連確率1 1 4 3、出現回数1 1 4 4、抽出元テキスト1 1 4 5）に設定したレコードを生成して単語テーブル1 1 4に登録する（S 1 4 1 8）。

[0084] 続いて、単語抽出部1 3 0は、S 1 4 1 3で選択中のレコードの抽出元テキスト1 1 4 5に格納されているテキストデータから抽出した全ての単語を単語Wとして選択済か否かを判定する（S 1 4 1 9）。全ての単語を選択済でなければ（S 1 4 1 9：NO）、処理はS 1 4 1 2に戻り、未選択の単語を単語Wとして選択して以上の同様の処理（S 1 4 1 4～S 1 4 1 8の処理を行う。一方、抽出した全ての単語を単語Wとして選択済であれば（S 1 4 1 9：YES）、処理はS 1 4 2 0に進む。

[0085] S 1 4 2 0では、単語抽出部1 3 0は、S 1 4 1 2で文書情報テーブル1 1 1の全てのレコードを選択済か否かを判定する。全てのレコードを選択済でなければ（S 1 4 2 0：NO）、処理はS 1 4 1 2に戻り、未選択のレコードを選択して以上と同様の処理S 1 4 1 3～S 1 4 1 8の処理を行う。一方、全ての単語を選択済であれば（S 1 4 2 0：YES）、単語抽出処理S 1 3 1 1は終了し、処理は類義語判定処理S 1 3 0 0の次のステップ（類義語候補生成処理S 1 3 1 2）に進む。

[0086] 図1 5は、図1 3に示した類義語候補生成処理S 1 3 1 2の詳細を説明するフローチャートである。以下、同図とともに類義語候補生成処理S 1 3 1 2について説明する。

[0087] まず、類義語候補生成部1 4 0が、単語テーブル1 1 4を取得する（S 1 5 1 1）。

[0088] 続いて、類義語候補生成部1 4 0は、同じカテゴリに属する2つの単語（単語A、単語B）を単語テーブル1 1 4から選択する（S 1 5 1 2）。

[0089] 続いて、類義語候補生成部1 4 0は、選択した単語Aと単語Bの夫々のカテゴリ関連確率を単語テーブル1 1 4から取得する（S 1 5 1 3）。

- [0090] 続いて、類義語候補生成部140は、文書情報テーブル111に基づき単語Aと単語Bの共起度数（類似度）を求める（S1514）。尚、共起度数の算出方法は必ずしも限定されないが、例えば、公知の各種機械学習手法（深層学習（DNN（Deep Neural Network）、RNN（Recurrent Neural Network）等）を用いて求める。
- [0091] 続いて、類義語候補生成部140は、単語Aと単語Bとの間の編集距離を求める。尚、類義語候補生成部140は、例えば、編集距離を、単語Aと単語Bの長さの和を用いて正規化する（S1515）。
- [0092] 続いて、類義語候補生成部140は、単語A、単語B、正誤情報（＝「不明」）、単語Aと単語Bの夫々が属するカテゴリ、S1514で求めた共起度数、S1515で求めた編集距離、単語Aと単語Bの夫々のカテゴリ関連確率、単語テーブル114から取得される単語Aと単語Bの夫々の出現回数、および単語Aと単語Bの夫々の抽出元の文章（テキストデータ）を、夫々対応する項目（単語A1151、単語B1152、正誤情報1153、単語カテゴリ1154、共起度数1155、編集距離1156、カテゴリ関連確率1157、出現回数1158、抽出元テキスト1159）に格納したレコードを生成して類義語候補テーブル115に登録する（S1516）。
- [0093] 続いて、類義語候補生成部140は、単語テーブル114から2つの単語の組合せを全て選択済か否かを判定する（S1517）。全ての組合せを選択済でなければ（S1517：NO）、処理はS1512に戻り、未選択の組合せについて以上と同様の処理を行う。一方、全ての組合せを選択済であれば（S1517：YES）、類義語候補生成処理S1312は終了し、処理は類義語判定処理S1300の次のステップ（類義語抽出ルール適用処理S1313）に進む。
- [0094] 図16は、図13に示した類義語抽出ルール適用処理S1313の詳細を説明するフローチャートである。以下、同図とともに類義語抽出ルール適用処理S1313について説明する。
- [0095] まず、類義語抽出ルール適用部150が、類義語候補テーブル115を取

得する（S 1 6 1 1）。

[0096] 続いて、類義語抽出ルール適用部 1 5 0 は、類義語抽出ルール（閾値テーブル 1 1 7、部分列正誤テーブル 1 1 8）を取得する（S 1 6 1 2）。

[0097] 続いて、類義語抽出ルール適用部 1 5 0 は、類義語候補テーブル 1 1 5 のレコードを一つ選択する（S 1 6 1 3）。

[0098] 続いて、類義語抽出ルール適用部 1 5 0 は、選択中のレコードにおける関係性特徴量を閾値テーブル 1 1 7 の閾値と比較し、閾値未満のものがあるか否かを判定する（S 1 6 1 4）。具体的には、類義語抽出ルール適用部 1 5 0 は、当該レコードにおける関係性特徴量（共起度数 1 1 5 5、編集距離 1 1 5 6、カテゴリ関連確率 1 1 5 7 における単語 A と単語 B の夫々のカテゴリ関連確率、出現回数 1 1 5 8 における単語 A と単語 B の夫々の出現回数）のうち、対応する閾値テーブル 1 1 7 の閾値未満のものがないか判定する。尚、上記判定において、類義語抽出ルール適用部 1 5 0 は、単語 A、単語 B のカテゴリ関連確率の閾値については、単語 A と単語 B が属する共通のカテゴリの閾値テーブル 1 1 7 のカテゴリ関連確率閾値 1 1 7 2 に格納されている値を用いる。閾値未満の関係性特徴量が一つでもあった場合（S 1 6 1 4 : Y E S）、類義語抽出ルール適用部 1 5 0 は、選択中のレコードの単語ペアを非類義語辞書 1 2 2 に登録し（S 1 6 2 1）、選択中のレコードを類義語候補テーブル 1 1 5 から削除する（S 1 6 2 2）。その後、処理は S 1 6 2 0 に進む。一方、閾値未満の関係性特徴量が一つもなかった場合（S 1 6 2 1 : N O）、処理は S 1 6 1 5 に進む。

[0099] S 1 6 1 5 では、類義語抽出ルール適用部 1 5 0 は、選択中のレコードの単語 A と単語 B を対照することにより不一致部分列を特定し、特定した不一致部分列の組合せ（不一致部分列ペア）を不一致部分列テーブル 1 1 6 に格納する処理（以下、「不一致部分列特定処理 S 1 6 1 5」と称する。）を行う。不一致部分列特定処理 S 1 6 1 5 の詳細については後述する。

[0100] 続いて、類義語抽出ルール適用部 1 5 0 は、不一致部分列テーブル 1 1 6 からレコードを 1 つ選択する（S 1 6 1 6）。以下、S 1 6 1 6 で選択した

レコードを選択中の部分列レコードと称する。

[0101] 続いて、類義語抽出ルール適用部150は、選択中の部分列レコードの不一致部分列ペアに対応する部分列正誤テーブル118の正誤情報1184の値を取得し、取得した値が「誤」であるか否かを判定する(S1617)。取得した値が「誤」であった場合(S1617: YES)、類義語抽出ルール適用部150は、S1613で選択したレコードの単語ペアを非類義語辞書122に登録し(S1621)、選択中のレコードを類義語候補テーブル115から削除する(S1622)。その後、処理はS1620に進む。一方、取得した値が「正」であった場合もしくは正誤情報1184に値が未設定である場合(S1617: NO)、処理はS1618に進む。

[0102] S1618では、類義語抽出ルール適用部150は、S1617で取得した値が「正」であるか否かを判定する。取得した値が「正」であった場合(S1618: YES)、類義語抽出ルール適用部150は、S1613で選択したレコードの単語ペアを類義語辞書に登録し(S1620)、選択中のレコードを類義語候補テーブル115から削除する(S1622)。その後、処理はS1620に進む。一方、取得した値が「正」でなかった場合(S1618: NO)、処理はS1619に進む。

[0103] S1619では、類義語抽出ルール適用部150は、S1616において不一致部分列テーブル116の全てのレコードを選択済か否かを判定する。全てのレコードを選択済でなければ(S1619: NO)、処理はS1617に戻り、未選択のレコードを選択して以上と同様の処理を行う。一方、全てのレコードを選択済であれば(S1619: YES)、処理はS1620に進む。

[0104] S1620では、類義語抽出ルール適用部150は、S1613において類義語候補テーブル115から全てのレコードを選択済か否かを判定する。全てのレコードを選択済でなければ(S1620: NO)、処理はS1613に戻り、次のレコードを選択して以上と同様の処理を行う。一方、全てのレコードを選択済であれば(S1620: YES)、類義語抽出ルール適用

処理 S 1 3 1 3 は終了し、処理は類義語判定処理 S 1 3 0 0 の次のステップ（類義語候補正誤判定処理 S 1 3 1 4）に進む。

[0105] 図 1 7 は、図 1 6 に示した不一致部分列特定処理 S 1 6 1 5 の詳細を説明するフローチャートである。以下、同図とともに不一致部分列特定処理 S 1 6 1 5 について説明する。

[0106] まず、類義語抽出ルール適用部 1 5 0 が、単語 A と単語 B の間で一致している文字列の中で最大長のもの（以下、「一致部分列」と称する。）を取得する（S 1 7 1 1）。例えば、単語 A が「S I P ノズル」であり、単語 B が「真空ノズル」である場合、類義語抽出ルール適用部 1 5 0 は「ノズル」を一致部分列として取得する。

[0107] 続いて、類義語抽出ルール適用部 1 5 0 は、一致部分列の長さが 1 以下であるか否かを判定する（S 1 7 1 2）。一致部分列の長さが 1 以下である場合（S 1 7 1 2 : Y E S）、類義語抽出ルール適用部 1 5 0 は、空の（値を設定しない）不一致部分列テーブル 1 1 6 を生成して不一致部分列特定処理 S 1 6 1 5 を終了する（S 1 7 1 9）。

[0108] 一方、一致部分列の長さが 1 よりも大きい場合（S 1 7 1 2 : N O）、類義語抽出ルール適用部 1 5 0 は、単語 A と単語 B の夫々について、一致部分列の左側に存在する全ての文字列を夫々、単語 A と単語 B の左側不一致部分列として取得する（S 1 7 1 3）。例えば、単語 A が「S I P ノズル」、単語 B が「真空ノズル」である場合、一致部分列は「ノズル」、単語 A の左側不一致部分列は「S I P」、単語 B の左側不一致部分列は「真空」となる。

[0109] 続いて、類義語抽出ルール適用部 1 5 0 は、単語 A の左側不一致部分列もしくは単語 B の左側不一致部分列の長さが 1 以下であるか否かを判定する（S 1 7 1 4）。いずれかの単語の左側不一致部分列の長さが 1 以下である場合（S 1 7 1 4 : Y E S）、処理は S 1 7 1 6 に進む。

[0110] 一方、単語 A と単語 B のいずれも左側不一致部分列の長さが 2 以上である場合（S 1 7 1 4 : N O）、処理は S 1 7 1 5 に進み、類義語抽出ルール適用部 1 5 0 は、単語 A の左側不一致部分列と単語 B の左側不一致部分列を設

定したレコードを有する不一致部分列テーブル 116 を生成する。

[0111] S1716では、類義語抽出ルール適用部150は、単語A、単語Bの夫々について、一致部分列の右側に存在する全ての文字列を夫々、単語A、単語Bの右側不一致部分列として取得する。例えば、単語Aが「リンスチューブ」、単語Bが「リンスノズル」である場合、一致部分列は「リンス」、単語Aの右側不一致部分列は「ノズル」、単語Bの右側不一致部分列は「チューブ」となる。

[0112] 続いて、類義語抽出ルール適用部150は、単語Aの右側不一致部分列もしくは単語Bの右側不一致部分列の長さが1以下であるか否かを判定する（S1717）。いずれかの単語の右側不一致部分列の長さが1以下である場合（S1717：YES）、類義語抽出ルール適用部150は、空の（値を設定しない）不一致部分列テーブル116を生成して不一致部分列特定処理S1615を終了する（S1719）。

[0113] 一方、単語Aと単語Bのいずれも右側不一致部分列の長さが2以上である場合（S1717：NO）、類義語抽出ルール適用部150は、単語Aの右側不一致部分列と単語Bの右側不一致部分列を設定したレコードを有する不一致部分列テーブル116を生成し（S1718）、不一致部分列特定処理S1615は終了し、処理は呼び出し元の次のステップに進む。

[0114] 図18は、図13に示した類義語候補正誤判定処理S1314の詳細を説明するフローチャートである。以下、同図とともに類義語候補正誤判定処理S1314について説明する。

[0115] まず、類義語候補正誤判定部160が、類義語候補テーブル115を取得する（S1811）。

[0116] 続いて、類義語候補正誤判定部160は、ユーザ装置2を介して、類義語候補テーブル115に格納されている単語ペアについて正誤情報の入力をユーザから受け付ける（S1812）。ユーザ装置2は、例えば、類義語候補テーブル115の単語ペアの一覧が記載されるとともに、単語ペアの正誤情報の入力欄を有する画面（以下、「正誤判定入力画面1900」と称する。

)を表示してユーザから各単語ペアの正誤情報の入力を受け付ける。

[0117] 図19に、正誤判定入力画面1900の一例を示す。例示する正誤判定入力画面1900は、類義語候補となる単語ペアの総数の表示欄1901、類義語候補テーブル115に記載された単語ペアの表示欄(単語Aの表示欄1911、単語Bの表示欄1912)、正誤情報入力欄1920、単語ペアの抽出元の文書(テキストデータ)の表示欄1930、および正誤判定登録ボタン1940を有する。

[0118] ユーザは、例えば、単語ペアの抽出元の文書(テキストデータ)の表示欄1930の内容を参考にしつつ、正誤情報入力欄1920に表示されているチェックボックスを操作することにより各単語ペアの正誤情報を入力する。尚、ユーザは、正誤判定入力画面1900に表示されている全ての類義語候補(単語ペア)について正誤情報を入力しなくてもよい。全ての類義語候補について正誤情報を入力しない場合でも、入力された正誤情報を利用して類義語抽出ルールが生成され、正誤情報が入力されなかった類義語候補についても類義語抽出ルール適用処理S1313において類義語か否かについて判定が行われる。

[0119] 図18に戻り、続いて、類義語候補正誤判定部160は、ユーザから正誤情報を受け付けているレコード(正誤情報1153が格納されているレコード)を類義語候補テーブル115から1つ選択する(S1813)。

[0120] 続いて、類義語候補正誤判定部160は、選択中のレコードの正誤情報1153に格納されている値を確認する(S1814)。正誤情報1153に「正」が格納されている場合(S1814:正)、類義語候補正誤判定部160は、当該レコードの類義語候補(単語ペア)を類義語辞書121に登録する(S1815)。その後、処理はS1819に進む。一方、選択中のレコードの正誤情報1153に「誤」が格納されている場合(S1814:誤)、類義語候補正誤判定部160は、当該単語ペアを非類義語辞書122に登録する(S1817)。その後、処理はS1819に進む。

[0121] S1819では、類義語候補正誤判定部160は、当該レコードを類義語

候補テーブル 115 から削除する (S1819)。

[0122] S1820では、類義語候補正誤判定部160は、S1813で、正誤情報を受け付けている類義語候補(単語ペア)を類義語候補テーブル115から全て選択済みかを判定する(S1820)。全ての単語ペアを選択済みであれば(S1820:NO)、処理はS1813に戻り、類義語候補正誤判定部160は、次の類義語候補(単語ペア)について処理を行う。一方、全ての類義語候補(単語ペア)を選択済みであれば(S1820:YES)、類義語候補正誤判定処理S1314は終了し、処理は類義語判定処理S1300の次のステップ(類義語抽出ルール生成処理S1315)に進む。

[0123] 図20は、図13の類義語抽出ルール生成処理S1315の詳細を説明するフローチャートである。同図示のように、類義語抽出ルール生成処理S1315は、閾値決定処理2010と部分列正誤テーブル生成処理S2020を含む。

[0124] 図21は、図20に示した閾値決定処理S2010の詳細を説明するフローチャートである。閾値決定処理2010では、閾値決定部181が、類義語候補テーブル115に基づき閾値テーブル117を更新する。以下、同図とともに閾値決定処理S2010について説明する。

まず、閾値決定部181は、類義語候補テーブル115を取得する(S2111)。

[0125] 続いて、閾値決定部181は、単語カテゴリリスト112を取得する(S2112)。

[0126] 続いて、閾値決定部181は、単語カテゴリリスト112からカテゴリを一つ選択する(S2113)。

[0127] 続いて、閾値決定部181は、選択中のカテゴリについて、関係性特徴量ごとに、類義語候補テーブル115に基づき、正誤情報1153に「正」が格納されている単語ペアの数(以下、「正の数」と称する。)と、正誤情報1153に「誤」が格納されている単語ペアの数(以下、「誤の数」と称する。)の関係性特徴量の値に応じた分布である特徴量正誤数分布を生成する。

(S 2 1 1 4)。

[0128] 続いて、閾値決定部 1 8 1 は、各関係性特徴量について、夫々の特徴量正誤数分布に基づき、「正の数」と「誤の数」との差の符号が反転する値を特定し、特定した値に基づき夫々の閾値を設定する (S 2 1 1 5)。

[0129] 図 2 2 に、ある関係性特徴量の特徴量正誤数分布の一例を示す。この例の場合、関係性特徴量の値が「0. 7 5 - 0. 8」と「0. 8 - 0. 8 5」の間で正の数と誤の数の差が反転しているので、閾値決定部 1 8 1 は、例えば、当該関係性特徴量の閾値を「0. 8 2」に設定する。尚、上記のような閾値の設定方法は一例に過ぎず、他の方法（例えば、上記反転する部分から所定のマージンを設ける、精度と再現率の積が最大となる値を閾値とする等）で設定してもよい。また例えば、ユーザ装置 2 を介してユーザから閾値の設定を受け付けるようにしてもよい。また、その場合、図 2 2 に示すようなグラフをユーザに提示するようにしてもよい。

[0130] 図 2 1 に戻り、続いて、閾値決定部 1 8 1 は、選択中のカテゴリを単語カテゴリ 1 1 7 1 に、S 2 1 1 5 で決定した各関係性特徴量の閾値を対応する項目（カテゴリ関連確率閾値 1 1 7 2、出現回数閾値 1 1 7 3、共起度数閾値 1 1 7 4、編集距離閾値 1 1 7 5）に格納したレコードを閾値テーブル 1 1 7 に登録する (S 2 1 1 6)。

[0131] 続いて、閾値決定部 1 8 1 は、S 2 1 1 3 で単語カテゴリリスト 1 1 2 の全てのカテゴリを選択済か否かを判定する (S 2 1 1 7)。未選択のカテゴリがある場合 (S 2 1 1 3 : NO)、処理は S 2 1 1 3 に戻り、未選択のカテゴリについて以上と同様の処理を行う。全てのカテゴリを選択済であれば (S 2 1 1 7 : YES)、閾値決定処理 S 2 0 1 0 は終了し、処理は類義語抽出ルール生成処理 S 1 3 1 5 の次のステップ（部分列正誤テーブル生成処理 S 2 0 2 0）に進む。

[0132] 図 2 3 は、図 2 0 に示した部分列正誤テーブル生成処理 S 2 0 2 0 の詳細を説明するフローチャートである。部分列正誤テーブル生成処理 S 2 0 2 0 では、類義語抽出ルール生成部 1 8 0 の部分列正誤テーブル生成部 1 8 2 が

、類義語候補テーブル 115 の正誤情報 1153 の値に基づき、類義語候補テーブル 115 の類義語候補（単語ペア）の不一致部分列に類似性があるか否かを判定し、判定した結果を部分列正誤テーブル 118 の正誤情報 1184 に格納する。以下、同図とともに部分列正誤テーブル生成処理 S2020 について説明する。

- [0133] まず、部分列正誤テーブル生成部 182 は、類義語候補テーブル 115 を取得する（S2311）。
- [0134] 続いて、部分列正誤テーブル生成部 182 は、類義語候補テーブル 115 からレコードを 1 つ選択する（S2312）。
- [0135] 続いて、部分列正誤テーブル生成部 182 は、S2312 で選択したレコードから、類義語候補（単語ペア：単語 A 1151、単語 B 1152）、正誤情報 1153、および単語カテゴリ 1154 の夫々に格納されている値を取得する（S2313）。
- [0136] 続いて、部分列正誤テーブル生成部 182 は、選択中のレコードの類義語候補（単語ペア）に一致部分列があるか否かを判定する（S2314）。例えば、単語 A が「SIP ノズル」、単語 B が「真空 ノズル」である場合、部分列「ノズル」が一致しているため、部分列正誤テーブル生成部 182 は、当該類義語候補（単語ペア）は一致部分列があると判定とする。
- [0137] 類義語候補（単語ペア）の一致部分列がない場合（S2314：NO）、処理は S2319 に進む。一方、一致部分列がある場合（S2324：YES）、部分列正誤テーブル生成部 182 は、図 17 に示した不一致部分列特定処理 S1615 を実行して不一致部分列を特定し、特定した不一致部分列を内容とする不一致部分列テーブル 116 を生成する。
- [0138] 続いて、部分列正誤テーブル生成部 182 は、S1615 で生成した不一致部分列テーブル 116 からレコードを 1 つ選択する（S2315）。
- [0139] 続いて、部分列正誤テーブル生成部 182 は、選択したレコードの単語 A と単語 B の不一致部分列（部分列 A 1161、部分列 B 1162）、S2323 で取得した正誤情報 1153 および単語カテゴリ 1154 の各内容に対

応する各項目（部分列A 1 1 8 2、部分列B 1 1 8 3、正誤情報 1 1 8 4、単語カテゴリ 1 1 8 1）に格納したレコードを生成して部分列正誤テーブル 1 1 8 に登録する（S 2 3 1 6）。

[0140] 続いて、部分列正誤テーブル生成部 1 8 2 は、S 2 3 1 5 において不一致部分列テーブル 1 1 6 の全てのレコードを選択済か否かを判定する（S 2 3 1 8）。全てのレコードを選択済でなければ（S 2 3 1 8 : NO）、処理は S 2 3 1 5 に戻り、未選択のレコードを選択して以上と同様の処理を行う。一方、全てのレコードを選択済であれば（S 2 3 1 8 : YES）、処理は S 2 3 1 9 に進む。

[0141] S 2 3 1 9 では、部分列正誤テーブル生成部 1 8 2 は、S 2 3 1 2 で類義語候補テーブル 1 1 5 のレコードを全て選択済か否かを判定する。レコードを全て選択済でなければ（S 2 3 1 9 : NO）、処理は S 2 3 1 2 に戻り、未選択のレコードを選択して以上と同様の処理を行う。一方、レコードを全て選択済であれば（S 2 3 1 9 : YES）、部分列正誤テーブル生成処理 S 2 0 2 0 は終了する。また、類義語抽出ルール生成処理 S 1 3 1 5 が終了し、処理は類義語判定処理 S 1 3 0 0 の次のステップ（S 1 3 1 6）に進む。

[0142] 図 1 3 に戻り、続いて、類義語判定装置 1 0 0 は、類義語候補テーブル 1 1 5 を再提示することについての要求の有無をユーザから受け付ける（S 1 3 1 6）。この受け付けは、例えば、図 1 9 に示した正誤判定入力画面 1 9 0 0 において、ユーザが正誤判定登録ボタン 1 9 4 0 を操作することによりユーザ装置 2 に表示される画面（以下、「再提示要否確認ダイアログ 1 9 6 0」と称する。）を介して行われる。ユーザは、例えば、類義語抽出ルール適用処理 S 1 3 1 3 により類義語抽出ルールが類義語候補テーブル 1 1 5 に適用されることにより更新された後の類義語候補テーブル 1 1 5 の類義語候補（単語ペア）に正誤情報を設定しようとする際に再提示の要求を行う。

[0143] 図 2 4 に再提示要否確認ダイアログ 1 9 6 0 の一例を示す。同図に示すように、例示する再提示要否確認ダイアログ 1 9 6 0 には、「はい」ボタン 1 9 6 1 と「いいえ」ボタン 1 9 6 2 が設けられている。再提示を要求する場

合、ユーザは「はい」ボタン１９６１を操作する。また、再提示を要求しない場合、ユーザは「いいえ」ボタン１９６２を操作する。

[0144] 図１３に戻り、ユーザが「はい」ボタン１９６１を操作すると（Ｓ１３１６：ＹＥＳ）、処理はＳ１３１３に戻り、類義語判定装置１００は、ユーザから受け付けた正誤情報に基づき更新された類義語抽出ルール（閾値テーブル１１７、部分列正誤テーブル１１８）を類義語候補テーブル１１５に適用して、Ｓ１３１４からの処理を再度実行する。その際、類義語判定装置１００は、更新後の類義語抽出ルールを適用することにより更新された後の類義語候補テーブル１１５の単語ペアの一覧を記載した正誤判定入力画面１９００をユーザ装置２を介してユーザに提示し正誤判定の入力を再度受け付ける。尚、再提示要否確認ダイアログ１９６０において、ユーザが「いいえ」ボタン１９６２を操作した場合は（Ｓ１３１６：ＮＯ）、類義語判定処理Ｓ１３００は終了する。

[0145] このように、Ｓ１３１３～Ｓ１３１６の処理が繰り返し実行されることにより、類義語抽出ルールが更新され、当該類義語抽出ルールが類義語候補テーブル１１５に適用されることにより、類義語が類義語辞書１２１に、非類義語が非類義語辞書１２２に自動的に登録されていく。

[0146] 以上に説明したように、本実施形態の類義語判定システム１は、類義語候補の一部に対する正誤情報に基づき類義語抽出ルール（閾値テーブル１１７、部分列正誤テーブル１１８）を生成し、生成した類義語抽出ルールを他の類義語候補に適用することにより文書データから類義語や非類義語を抽出する。そのため、ユーザは少ない負荷で効率よく類義語辞書１２１および非類義語辞書１２２を作成することができる。また、類義語判定システム１は、類義語抽出ルール（閾値テーブル１１７、部分列正誤テーブル１１８）は、ユーザが入力した正誤情報に基づき（人が判断した情報を用いて）生成するので、文書データが少ない場合でも精度よく義語や非類義語を抽出することができる。

[0147] 以上、本発明の一実施形態について説明したが、本発明は上記の実施形態

に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。例えば、上記の実施形態は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、上記実施形態の構成の一部について、他の構成の追加や削除、置換をすることが可能である。

[0148] また、上記の各構成、機能部、処理部、処理手段等は、それらの一部または全部を、例えば、集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリやハードディスク、SSD (Solid State Drive) 等の記録装置、ICカード、SDカード、DVD等の記録媒体に置くことができる。

[0149] また、以上に説明した各情報処理装置の各種機能部、各種処理部、各種データベースの配置形態は一例に過ぎない。各種機能部、各種処理部、各種データベースの配置形態は、これらの装置が備えるハードウェアやソフトウェアの性能、処理効率、通信効率等の観点から最適な配置形態に変更し得る。

[0150] また、前述した各種のデータを格納するデータベースの構成（スキーマ (Schema) 等）は、リソースの効率的な利用、処理効率向上、アクセス効率向上、検索効率向上等の観点から柔軟に変更し得る。

符号の説明

[0151] 1 類義語判定システム、2 ユーザ装置、4 データ管理装置、110 記憶部、111 文書情報テーブル、112 単語カテゴリリスト、113 単語カテゴリ判定モデル、114 単語テーブル、115 類義語候補テーブル、116 不一致部分列テーブル、117 閾値テーブル、118 部分列正誤テーブル、121 類義語辞書、122 非類義語辞書、130 単語抽出処理部、140 類義語候補生成部、150 類義語抽出ルール適用部、160 類義語候補正誤判定部、170 不一致部分列特定部、180 類義語抽出ルール生成部、181 閾値決定部、182 部分列正誤

テーブル生成部、１９００ 正誤判定入力画面、１９６０ 再提示要否確認
ダイアログ、Ｓ１３００ 類義語判定処理、Ｓ１３１１ 単語抽出処理、Ｓ
１３１２ 類義語候補生成処理、Ｓ１３１３ 類義語抽出ルール適用処理、
Ｓ１３１４ 類義語候補正誤判定処理、Ｓ１３１５ 類義語抽出ルール生成
処理、Ｓ２０１０ 閾値決定処理、Ｓ２０２０ 部分列正誤テーブル生成処
理

請求の範囲

- [請求項1] プロセッサおよびメモリを有する情報処理装置を用いて構成され、
 文書データから抽出される複数の単語から選択される2つの単語の
 組合せである複数の類義語候補の一部について、構成する2つの単語
 が類義語の関係となるか否かを示す情報である正誤情報を取得し、
 前記文書データから取得される前記類義語候補についての特徴量と
 前記正誤情報とに基づき、前記類義語候補を構成する2つの単語が類
 義語の関係となるか否かを判定するための情報である類義語抽出ル
 ルを生成し、
 前記正誤情報を取得していない前記類義語候補に前記類義語抽出ル
 ルを適用することにより、構成する2つの単語が類義語の関係とな
 る前記類義語候補を抽出する、
 類義語判定システム。
- [請求項2] 請求項1に記載の類義語判定システムであって、
 前記正誤情報は、ユーザインタフェースを介してユーザから受け付
 けることにより取得した情報である、
 類義語判定システム。
- [請求項3] 請求項2に記載の類義語判定システムであって、
 前記正誤情報を取得していない前記類義語候補から、前記類義語抽
 出ルールを適用することにより抽出した前記類義語候補を除外した後
 の前記類義語候補をユーザに提示し、前記除外した後の類義語候補に
 についての前記正誤情報を受け付ける、
 類義語判定システム。
- [請求項4] 請求項3に記載の類義語判定システムであって、
 前記除外した後の類義語候補の提示を指示するための画面をユーザ
 に提示し、前記画面を介して前記指示を受け付けた場合に、前記除外
 した後の類義語候補の提示と前記類義語候補を除外した後の前記類義
 語候補についての前記正誤情報を受け付ける、

類義語判定システム。

[請求項5]

請求項 1 に記載の類義語判定システムであって、
前記文書データから抽出される前記複数の単語をカテゴリに分類し、
同じ前記カテゴリに属する 2 つの単語の組合せを前記類義語候補とする、

類義語判定システム。

[請求項6]

請求項 1 に記載の類義語判定システムであって、
前記特徴量は、前記類義語候補を構成する 2 つの単語の共起度数、
前記 2 つの単語の編集距離、前記 2 つの単語の夫々の抽出元の前記文書データにおける出現頻度のうちの少なくともいずれかである、
類義語判定システム。

[請求項7]

請求項 5 に記載の類義語判定システムであって、
前記特徴量は、前記類義語候補を構成する 2 つの単語の夫々の前記カテゴリとの関連の度合いを示す情報である、
類義語判定システム。

[請求項8]

請求項 1 に記載の類義語判定システムであって、
前記類義語候補に前記類義語抽出ルールを適用することにより、構成する 2 つの単語が非類義語の関係となる前記類義語候補を抽出する、
類義語判定システム。

[請求項9]

請求項 1 に記載の類義語判定システムであって、
前記正誤情報を取得している前記類義語候補について、構成する 2 つの単語の文字列が一致しない部分である不一致部分列を特定し、
2 つの単語の夫々の前記不一致部分列と前記正誤情報とを対応づけた情報である部分列正誤情報を前記類義語抽出ルールとして生成し、
前記類義語候補を構成する 2 つの単語について前記不一致部分列を特定し、特定した前記不一致部分列に前記類義語抽出ルールを適用す

ることにより、構成する２つの単語が類義語の関係となる前記類義語候補を抽出する、

類義語判定システム。

[請求項10]

請求項５に記載の類義語判定システムであって、

前記正誤情報を取得している前記類義語候補について、構成する２つの単語の文字列が一致しない部分である不一致部分列を特定し、

２つの単語の夫々の前記不一致部分列、前記２つの単語が属する前記カテゴリ、および前記正誤情報を対応づけた情報である部分列正誤情報を前記類義語抽出ルールとして生成し、

前記類義語候補を構成する２つの単語について前記不一致部分列を特定し、前記２つの単語が属する前記カテゴリの前記部分列正誤情報を前記類義語抽出ルールとして適用することにより、構成する２つの単語が類義語の関係となる前記類義語候補を抽出する、

類義語判定システム。

[請求項11]

請求項１に記載の類義語判定システムであって、

前記正誤情報が取得されている前記類義語候補について、前記特徴量の値に対する、前記正誤情報が正の前記類義語候補の数と前記正誤情報が誤の前記類義語候補の数の分布である特徴量正誤数分布を求め、

前記特徴量正誤数分布に基づき前記特徴量の閾値を決定し、

決定した前記閾値を前記類義語抽出ルールとする、

類義語判定システム。

[請求項12]

請求項１１に記載の類義語判定システムであって、

前記特徴量正誤数分布と前記閾値を示す画面をユーザに提示する、
類義語判定システム。

[請求項13]

請求項１に記載の類義語判定システムであって、

類義語辞書を記憶し、

２つの単語が類義語の関係となるとして抽出した前記類義語候補を

前記類義語辞書に登録する、
類義語判定システム。

- [請求項14] 請求項8に記載の類義語判定システムであって、
非類義語辞書を記憶し、
2つの単語が非類義語の関係となるとして抽出した前記類義語候補
を前記非類義語辞書に登録する、
類義語判定システム。
- [請求項15] プロセッサおよびメモリを有する情報処理装置が、
文書データから抽出される複数の単語から選択される2つの単語の
組合せである複数の類義語候補の一部について、構成する2つの単語
が類義語の関係となるか否かを示す情報である正誤情報を取得するス
テップと、
前記文書データから取得される前記類義語候補についての特徴量と
前記正誤情報とに基づき、前記類義語候補を構成する2つの単語が類
義語の関係となるか否かを判定するための情報である類義語抽出ルー
ルを生成するステップと、
前記正誤情報を取得していない前記類義語候補に前記類義語抽出ル
ールを適用することにより、構成する2つの単語が類義語の関係とな
る前記類義語候補を抽出するステップと、
を実行する、類義語判定方法。