

明 細 書

発明の名称：

同義関係判定装置、同義関係判定方法、及びそのプログラム

技術分野

[0001] 本発明は、同義関係期間判定装置、同義関係期間判定方法、及び、同義関係期間判定プログラムに関し、特に同義関係が成立する期間を判定する同義関係期間判定装置、同義関係期間判定方法、及び、同義関係期間判定プログラムに関する。

背景技術

[0002] 自然語データを対象とする情報処理では、自然語と自然語との意味上の関係をコンピュータで計算可能な状態で自動的に取得することが難しい。意味上の関係としては、上位概念と下位概念との関係や、同義語となる同義関係などがある。自然言語アプリケーションでは、特に、同義関係の獲得と利用が重要である。

[0003] 特に、インターネット上の風評の監視といったタスクの場合、組織名や商品名の正式名称を検索クエリとし、それにマッチした文書を監視対象とするのは不十分で、組織名や商品名の別名、省略形、伏字、隠語といった同義表現を獲得し、それらを検索クエリに追加することで、監視のものを少なくすることが必要となる。

そのため、同義表現を自動獲得するさまざまな手法が提案されている。

[0004] 例えば、非特許文献 1 では、同義表現候補の中で出現する文脈が類似しているものを同義表現として自動的に獲得しようとする手法が提案されている。すなわち、ある文に出現する語の共通性に基づいて、同時に使用される語が共通する際に、同義関係があると自動判定を試みている。

[0005] 又、特許文献 1 には、時間と共に同義関係が変化することに対応した同義語辞書の自動作成を目的として、各検索語の使用頻度の時系列の相関係数に基づいて単語同士の関連度を定義する手法が記載されている。

特許文献2には、「○菱電気」・「某A庁」等の隠語と「三菱電機」・「防衛庁」等の元の表現との同義関係を抽出することを目的として、「○」等の隠語表現リストから照合用索引を生成して、元の表現と照合することにより同義語関係を抽出する手法が開示されている。

- [0006] 特許文献3には、番組名とその略称や愛称等の同義関係を抽出することを目的として、放送局と放送時間の情報を使用して、同義語の候補からシリーズ名や各回タイトルを除いた語を同義語とする手法が開示されている。

先行技術文献

特許文献

- [0007] 特許文献1：特開平11-312168号公報
特許文献2：特開2003-296354号公報
特許文献3：特開2006-163710号公報

非特許文献

- [0008] 非特許文献1：寺田ら、“文脈情報による同義語辞書作成支援ツール”、社団法人情報処理学会研究報告 自然言語処理研究会報告2006（124）、2006年11月、p. 87-94

発明の概要

発明が解決しようとする課題

- [0009] しかしながら、同義表現の中でも省略語や伏字のなかには、曖昧性がある場合や、時間的に意味が変化する場合があるため、既存の同義表現獲得手法では正しく同義であると判定することが困難である。例えば、「東京電力」の伏字として「東○電力」が考えられるが、これは「東北電力」の伏字にもなりうる。このように、「東○電力」が指す内容は「東京電力」や「東北電力」等の曖昧性がある。
- [0010] また、実際には、時刻によって「東○電力」が指す内容が「東京電力」や「東北電力」に変化しうる。例えば、図13の例では、「東○電力」は、時刻A及び時刻Cでは「東京電力」を、時刻Bでは「東北電力」を指しており

、同義関係が時刻によって変化している例となっている。

[0011] このような場合、既存の方法では時刻によって内容が変化するような同義表現を考慮していないため正しく同義性を判定できない。同義性の判定に文脈を用いる非特許文献1のような手法では、時間情報を利用していないため時刻によって同義性が変化することを考慮していない。

[0012] また、時系列相関を用いる特許文献1のような手法でも、時刻によって同義性が変化する場合、図13に示すように「東○電力」は「東京電力」、「東北電力」のどちらとも相関が高くなり、同義関係があると判定できない。つまり、特許文献1の場合も時刻によって同義性が変化しないことを前提にして同義性を判定しているため、時刻によって同義性が変化する場合に同義性がある時区間を特定することができない。

[0013] 前述した非特許文献1では、文脈を用いて同義性を判定することができるが、時間情報を利用していないため、時刻で変化する同義性を把握することができない。

即ち、時間の推移によって同義性が変化し、一つの同義候補が時間によっては異なる同義元と同義となる場合には、特許文献1のような手法で算出する時系列相関は高くなり、その結果、同義関係を抽出できない。

[0014] 又、特許文献2記載の手法では、隠語や伏せ字に用いられる語（○、「ル」に対して「ノ」と「レ」の組み合わせ「ノレ」）を使用して同義元と同義となる同義候補を生成することができるが、同義候補の意味の時間変化をとらえることはできない。

特許文献3記載の手法では、同義語の判定に時間情報を使用しているが、同一の情報源（放送局）からの情報を対象としており、不特定多数から収集されるテキスト集合に対しては適用することができない。

[0015] 更に、上述した非特許文献1、特許文献1乃至3、及びこれらを組み合わせた技術では、同義候補の意味が時間とともに変化する際に、同義候補と同義元との同義性を正確に判定することができない、という不都合があった。

[0016] [発明の目的]

本発明は、不特定多数からのテキストで使用される自然語から、時間と共に意味が変化する同義候補の同義関係を有効に抽出し特定することを可能とした同義関係判定装置、同義関係判定方法、及びそのプログラムを提供することを、その目的とする。

課題を解決するための手段

[0017] 上記目的を達成するため、本発明に係る同義関係判定装置は、所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部と、外部入力されるテキスト中における前記同義表現候補と前記同義元表現との同義関係を一定の基準に基づいて判定し特定する同義関係判定特定手段とを備えている。

そして、前記同義関係判定特定部が、外部入力される前記テキストを収集しこれに基づいて発行時間が特定可能なテキスト集合を生成するテキスト収集部と、このテキスト収集部１４で収集されたテキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを特定し出力する同義候補検出手段と、前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する同義期間特定手段とを備えたことを特徴としている。

[0018] 上記目的を達成するため、本発明に係る同義関係判定方法は、所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部と、外部入力されるテキスト中における前記同義表現候補と同義元表現との同義関係を判定し特定する同義関係判定特定部を備えた同義関係判定装置にあって、

外部入力される前記テキストを収集しこれに基づいて発行時間が特定できるテキスト集合を、前記同義関係判定特定部のテキスト収集部が生成し（テキスト集合生成工程）、

この生成された前記テキスト集合に含まれる前記同義表現候補と前記同義

元表現との同義関係を、前記同義関係判定特定部が一定の基準に基づいて判定すると共に特定し（同義関係特定工程）、

前記同義関係を特定する工程にあっては、

前記テキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ前記同義関係判定特定部の同義候補検出手段が検索して特定し（同義候補検出工程）、

続いて、前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義関係判定特定部の同義期間特定手段が前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定するように構成したこと（同義期間特定工程）を特徴とする。

[0019] 上記目的を達成するため、本発明に係る同義関係判定用プログラムは、所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部10と、外部入力されるテキスト中における前記同義表現候補と同義元表現との同義関係を判定し特定する同義関係判定特定手段とを備えた同義関係判定装置にあって、

外部入力されるテキストを収集して発行時間を特定できるテキスト集合を生成するテキスト集合生成処理機能、および生成された前記テキスト集合に含まれる前記同義表現候補と前記同義元表現との同義関係を、一定の基準に基づいて判定すると共に特定処理する同義関係特定処理機能を設けると共に、

前記同義関係特定処理機能が、前記テキスト収集部で収集された前記テキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ検索して特定処理する同義候補検出処理機能、および前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する同義期間特定処理機能を含む構成とし、

これらの各処理機能を前記同義関係判定特定手段が備えているコンピュータに実現させるようにしたことを特徴としている。

発明の効果

[0020] 本発明は以上のように、同義表現候補が多く出現した時点を捉えて同義性の判定を行うように構成したので、同義関係が成立した開始時刻を出力することが可能となり、そのため、時刻によって同義性が変化する場合に、同義関係が成立する時区間を判定することが可能となるという前述した関連技術にない優れた同義関係判定装置、同義関係判定方法、及びそのプログラムを提供することができる。

図面の簡単な説明

[0021] [図1]本発明に係る同義関係判定装置の第1実施形態を示すブロック図である。

[図2]図1にて使用される同義候補リストの一例を示す説明図である。
[図3]図1にて使用される同義元の出現の時間推移と同義候補の出現の時間推移の例を示す説明図である。

[図4]図1に開示した第1実施形態の全体的な動作を示すフローチャートである。
[図5]図1に開示した第1実施形態における同義期間開始判定部の構成例を示すブロック図である。

[図6]図5に開示した同義期間開始判定部の他の構成例及び使用するテーブルの例を示す説明図である。

[図7]図6に示す同義期間開始判定部の他の構成例にあって、同義元の出現数を計算する処理の一例を示すフローチャートである。

[図8]同義元の出現数を使用して同義候補との同義期間の開始を判定する処理の一例を示すフローチャートである。

[図9]本発明に係る同義関係判定装置の第2実施形態を示すブロック図である。

[図10]図9に開示した第2実施形態の動作を示すフローチャートである。

[図11]本発明に係る同義関係判定装置の第3実施形態を示すブロック図である。

[図12]図11に開示した第3実施形態で使用する同義候補リストの一例を示す説明図である。

[図13]テキスト集合中の同義候補及び同義元の出現数の一例を示すグラフ図である。

[図14]第3実施形態における情報処理例を示すフローチャートである。

[図15]本発明の各実施形態に共通するハードウェア資源の構成例を示すブロック図である。

発明を実施するための形態

[0022] 以下、本発明の第1実施形態を図1乃至図8に基づいて説明する。

最初に、同義関係の概念を明確にすると共に本第1実施形態の基本的な構成内容を説明し、その後に本第1実施形態について、更に詳述する。

[0023] (同義関係の概念)

まず、本第1実施形態では、二つの語（自然語の語句）の同義関係を期間に対応させて判定するようにした。

ここで、同義関係は、種表現となる同義元と、この同義元と同義性を持つ可能性のある同義候補の表現とである。例えば、「日本電気」と「NEC」と「日電」という自然語の語句は、語としての同義性を持つ。そして、「日本電気」を種表現である同義元とすると、「NEC」「日電」は同義候補である。

[0024] そして、本第1実施形態では、時間的に変化する同義関係を抽出する。時間的に変化する同義関係としては、同義表現候補が多義的であって複数の同義元の候補となる場合や、同義元について時間的に関心や関係性が変化していくことで意味が変化する場合などがある。例えば「朝日」は多義的で、新聞、生命保険、飲料など異なる社名を同義元とする。そして、不特定多数から主にインターネットを介して発信されるテキスト（文章）では、「アサヒ」は朝には新聞を意味することが多く、夜には飲料を意味することが多い

という傾向が現れる可能性もある。

[0025] 同義元について時間的に関心や関係性が変化する例としては、番組中で成長する主人公の男優名が、「少年」、「青年」、「恋人」、「夫」、「父」、「祖父」などの語との同義関係を構築しては失うような時間変化がある。この例では、1年間の連続番組であれば数ヶ月毎に同義関係が変化する。

[0026] また、通常の辞書に使用されていた用語が流行語として別の意味を獲得することで同義関係を構築することがある。例えば、「サポーター」、「タマちゃん」、「埋蔵金」、「仕分け」、「なでしこ」などは、語の流行開始の前後で同義関係が変化する。流行語が特別な意味を獲得した後、その特別な意味を失うと、同義関係が再度変化する。このような変化は数年又は数十年単位となる。

[0027] 隠語や伏せ字を用いた表現（同義表現候補となる）は、多義的になりやすく、同義関係も変化しやすい。会社名をアルファベット一文字で表し、また、人物をアルファベット数文字のイニシャルで表すと、同義関係を構築可能な同義元は複数生じる。例えば「NEC」を「NE〇」と伏せ字表現すると、「NET」「NEW」「NEO」などの解釈も生じる。

社名をアルファベット表記している場合、何らかのニュースとの関係で伏せ字での電子的な対話がなされることがある。このような場合、ニュースとなったことで同義元の出現数が増加する時期に、当該伏せ字の出現が現れ、同義性を持つ。このようなニュース等による同義性は、数時間で終了することもある。

[0028] （基本的内容）

本第1実施形態にあって、同義関係判定装置101は、図1に示すように、所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部10と、外部入力されるテキスト中における前記同義表現候補と前記同義元表現との同義関係を一定の基準に基づいて判定し特定する同義関係判定特定手段12と、を備えている。

また、同義表現候補記録部10には、同義表現候補を生成するための種と

なる表現を入力しその種表現から同義候補を生成する同義候補生成部 10A が併設されている。

[0029] そして、上記同義関係判定特定手段 12 は、外部入力される前記テキストを収集しこれに基づいて発行時間が特定可能なテキスト集合を生成するテキスト収集部 14 と、このテキスト収集部 14 で収集されたテキスト集合から前述した同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを特定し出力する同義候補検出部 12A と、前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する同義期間特定部 12B とを備えている。

[0030] このため、上記基本構成によると、詳細は後述するが、同義表現候補が多く出現した時点を捉えて同義性の判定を行うようにしたので、同義関係が成立した開始時刻を出力することが可能となり、そのため、時刻によって同義性が変化する場合に、その同義関係が成立する時区間を判定することが可能となる。

[0031] ここで、上記同義候補検出部 12A は、前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を（前記同義表現候補が存在する時区間として）特定する同義候補検索部 16 と、同じく前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を（前記同義元表現が存在する時区間として）特定する同義元検索部 18 とを含む構成となっている。

[0032] 又、上記同義期間特定部 12B は、前記同義候補検出手段で前記同義表現候補が検出された時区間におけるテキスト集合中で少なくとも最も出現数の多い同義元表現を同義関係にある同義元として抽出する同義関係抽出部 20 と、この抽出された同義元表現が前記同義表現候補と同義関係にあると判定

すると共に当該両者が同義関係にある時区間の開始点を同義期間の開始点としてこれを同義期間と共に予め装備した同義辞書 3 2 に登録する同義期間開始判定部 2 2 とを含む構成となっている。

[0033] 上記した同義期間開始判定部 2 2 は、前記同義表現候補が前記テキスト集合で検出される時区間における集合中で、前記時区間での単位時間あたりの出現数と前記時区間以前での単位時間あたりの出現数の比率が最も大きい同義元表現についても、これを前記同義表現候補と同義関係にあると判定する機能を備えている。

[0034] 更に、前記同義期間特定部 1 2 B は、当該同義期間特定部 1 2 B の前記同義期間開始判定部 2 2 で同義関係にあると判定され特定された同義表現候補についてその単位期間あたりの出現数が予め設定したしきい値以下になった時点で、前記同義関係が解消した旨判定する同義期間終了判定部 2 4 を備えている。

[0035] ここで、上述した同義候補生成部 1 0 A は、前述したように同義表現候補を生成するための種となる表現（以下「種表現」と呼ぶ）を入力とし、その種表現から同義表現候補を生成する機能を備えている。この場合、同義表現候補は、種表現に対して部分文字列を抽出して省略語を生成する、種表現の一部を特定の文字に置換して伏字を生成する、種表現を別の言語に翻訳した翻訳表現を生成する、等の操作を複数回適用して、複数生成される。

[0036] また、同義表現候補記録部 1 0 は、同義候補生成部 1 0 A で生成した同義表現候補を記録する。同義表現候補を見出しとし、対応する種表現（以下、「同義元表現」と呼ぶ）を 1 個以上記録する。ここで、複数の同義元表現が記録された同義表現候補は、曖昧性のある同義表現候補である。

[0037] 同義候補検出部 1 2 A は、発行時間を特定できるテキスト集合を読んでいき、同義表現候補記録部 1 0 に記録されている同義表現候補、同義元表現が各時点で何回出現しているかを計数する機能を備えている。そして、同義表現候補の単位期間あたりの出現数が大きく増加している時区間を検出する。

[0038] 更に、同義期間判定手段 1 2 B は、同義候補検出部 1 2 A で検出した時区

間でのテキスト集合を用いて同義表現候補と同義関係にある同義元表現を特定し、それらが同義関係にある期間を同義辞書 32 に登録する。

[0039] また、同義期間判定手段 12B の一部を成す同義期間開始判定部 22 は、同義候補検出部 12A で検出された同義表現候補が、対応する同義元表現のどれと同義関係にあるかを判定し、同義関係であった場合、同義候補検出部 12A で検出された時区間の開始点を同義関係の開始点として同義辞書 32 に登録する。

[0040] この同義関係の判定は、同義表現候補が検出された時区間のテキスト集合中で、最も出現数の多い同義元表現、若しくは、前記時区間での単位時間あたりの出現数と前記時区間以前での単位時間あたりの出現数の比率が最も大きい同義元表現、を前記同義表現候補と同義関係にあると判定することで行う。

同義辞書 5 は、同義関係にある表現を記録する辞書であり、同義関係の開始、終了時刻もあわせて登録することができる。

[0041] 上述したように、本第 1 実施形態において、同義関係判定装置 101 は、同義表現候補記録部 10 と同義関係判定特定手段 12 とを備えている。そして、同義関係判定特定手段 12 は、テキスト収集部 14 と、同義候補検索部 16 と、同義元検索部 18 と、同義関係抽出部 20 と、同義期間開始判定部 22 とを備えている。更に、この同義関係判定装置 101 は、同義候補生成部 30 と、同義辞書 32 とを備えている。

そして、この構成により、同義関係が成立する時区間 PD を判定しようとするものである。

[0042] ここで、前述したように、同義候補検索部 16 と同義元検索部 18 とにより同義候補検出部 12A が構成され、同義関係抽出部 20 と同義期間開始判定部 22 とにより同義期間判定手段 12B が構成されている。

[0043] 同義表現候補記録部 10 は、同義元 OW である語に対して同義語となる候補の同義候補 EW を当該同義元 OW と関連させて予め記憶する。この同義表現候補記録部 10 は、同義候補 EW を見出しとし、対応する種表現（同義元

OW) を 1 個以上関係させて記憶する。 図 2 に示す同義候補リスト 10A は、このように同義候補 EW と同義元 OW とを関連させたデータである。同義候補リスト 10A は、データ収集の直前までに作成されれば良く、過去に収集したテキストを使用して同義候補 EW を自動生成しても良いし、ユーザが人為的に検討した同義候補 EW が入力されるようにしても良い。

[0044] 自動生成する例では、同義表現候補記録部 10 は、同義候補生成部 30 が自動生成した同義候補 EW を同義候補リスト 10A に含めて記憶する。ここで、複数の同義元 OW が記憶された同義候補 EW は、曖昧性のある同義候補 EW である。

[0045] 同義関係判定特定手段 12 は、例えばネットワーク 96 を介して情報処理可能な自然語データを含むテキストを収集すると共に、当該テキストの集合にデータ処理をする（図 1 参照）。このネットワーク 96 は、例えばインターネットであり、インターネットに接続されるローカルなネットワーク 96 でも良い。

[0046] 自然語データは、日本語、英語等の言語で、語、文、段落などを要素とする文章であり、人間が読み取るための情報である。テキストは、この自然語データを含むデータであり、自然語が表現されるものであればファイル形式はどのようなタイプでも良い。また、一行のコメントから、文書、論文や本としてまとめられたものまで、分量や校正の程度はどのようなものでも良い。

[0047] このテキストは、文章の著者と関連した論理的な所在と、発行時間とを属性情報として有すると良い。論理的な所在としては、IP アドレス、IP アドレス等で特定できるサーバ群（Web サイト）中のファイル所在やデータベースの検索結果をも示す URL などがある。

[0048] テキスト収集部 14 は、テキストを発行時間と関連させて収集することで、テキスト集合 TX を生成する。このテキスト収集部 14 は、発行時間（例えば、書き込み時間）を属性情報として有するテキストについては、当該発行時間に発行されたと扱い、発行時間が不明なテキストが新たに収集された

際には、収集した時間（例えば、クロール時間）を当該テキストの発行時間とすることができる。

[0049] このテキストの収集は、インターネットに接続された不特定多数のサーバ装置 70 へのロボット検索としても良いし、予めユーザによって所在が指定されたアドレスにアクセスすることで得られるデータを収集するようにしても良い。サーバ装置からテキストを収集する際には、文字データのみを収集するのではなく、画像やリンクを含むファイル全体を受信しても良い。既に収集したデータとの差分のみを受信するようにしても良い。

[0050] テキスト集合 TX は、大量の文章を含むテキストデータの集合で、発行時間又は発行時間が属する期間をキーとして同義表現候補記録部 10 等の記憶媒体に格納しておくとも良い。また、テキストのみではなく、同義候補リスト 10A にある同義元 OW 及び同義候補 EW へのインデックスを生成して一体的に格納しても良い。

[0051] 同義候補検索部 16 は、テキスト集合 TX に含まれる同義候補 EW が当該テキスト集合 TX で検索される時区間 PD を、当該テキストの発行時間から計算する。

時区間 PD は、開始時間で区切られる区間であり、本第 1 実施形態では、同義候補 EW が検索される期間である。収集したテキスト集合 TX から同義候補 EW が検索され始めた際には時区間 PD の開始時間はあるが継続中で有り終了時間はない。

[0052] 当該時区間 PD 中のテキスト集合 TX には当該同義候補 EW が一定数以上含まれ、時区間 PD 前のテキスト集合 TX には同義候補 EW が一定数以上は含まれない。一定数は、0 でも良いし、通常時（平時）の検索数としても良い。

[0053] 同義候補検索部 16 は、発行時間を特定できるテキスト集合 TX を読んでいき、同義表現候補記録部 10 に記憶されている同義候補 EW が各時点で何回出現しているかを計数する。そして、同義候補 EW が当該テキスト集合 TX で検索される期間を時区間 PD とする。例えば、同義候補 EW の単位期間

あたりの出現数が大きく増加している期間を当該同義候補EWの時区間PDとする。

[0054] 同義元検索部18は、同義候補EWが検索される時区間PDと重なる期間のテキスト集合TXから同義表現候補記録部10に記憶された同義元OWを検索することで、当該同義元OWの出現を計算する。「出現」は検索の結果得られるデータ項目であり、例えば出現数や出現比率である。

時区間PDと重なる期間は、時区間PDと完全に同一の期間でも良いし、時区間PD前の一定時間前からとしても良い。この重なる期間は、文字通り一部で時区間PDと重なれば良い。

[0055] 同義元検索部18は、時区間PDを有する同義候補EWと同義関係となる可能性のある同義元OWを、当該時区間PDと重なる期間のテキスト集合TXから検索する。すると、同義候補EWが出現した時区間PDと重なる期間（同一期間を含む）にて、同義候補EWがどのように出現したのかというデータを得ることができる。同義元OWについても時区間PDを計算するようにしても良い。

[0056] 同義元検索部18は、情報処理として、時区間PDとは独立して予め定められた検索指定時刻や単位時間毎に同義元OWを検索しておき、その検索結果から、さらに時区間PDと重なる期間のテキスト集合TXでの出現数を計算するようにしても良い。

いずれにせよ、同義元検索部18は、同義元OWが各時点で何回出現しているかを計数する。

[0057] そして、同義関係抽出部20は、同義候補EWが検索された時区間PDに同義元OWの出現がある際には、同義候補EWと同義元OWと間に同義関係を抽出する。例えば、時区間PDと同一の期間に出現がある同義元OWと、当該同義候補EWとの同義関係を抽出する。

[0058] また、時区間PDの前の期間の出現に対して時区間PD中の出現が増加している同義元OWと同義関係があると抽出しても良い。同義元OWの時区間PDを算出する例では、同義関係抽出部20は同義候補EWの時区間PDと

同義元OWの時区間PDとが重なる際に、同義候補EWの時区間PDに同義元OWが出現したと判定することができる。

[0059] このように、同義関係抽出部20は、同義候補検索部16及び同義元検索部18で検出した時区間PDでのテキスト集合TXを用いて同義候補EWと同義関係にある同義元OWを特定する。図1に示す同義辞書32と接続されている際には、この時区分PDで特定される同義関係を同義辞書32に登録する。

[0060] 図1に示す例では、同義候補EWが検索される時区間PDと重なる期間にテキスト集合TXに出現した同義元を検索することで、情報処理により同義候補EWと同義元OWとの同義関係を時区間PDで区切りつつ自動的に抽出することができる。

[0061] このような同義候補EWが検索される時区間PDで成立する同義関係を、「時区間同義」という。時区間同義が成立しうる期間は、上述のように様々な有りニュース等に起因する際には数時間で終了することもあり、一方流行語や新概念の登場などでは数十年成立し得る。また、そして、同義関係によっては、時区間同義の開始後、テキストTXの収集時点では終了していないこともある。

[0062] 同義候補生成部30は、同義候補EWを生成するための種となる種表現を入力とし、その種表現から同義候補EWを自動的に生成する。同義候補生成部30は、同義元OWの表現に対して、次の操作を複数回適用することで、複数の同義候補EWを自動生成する。

(1) 種表現に対して部分文字列を抽出して省略語を生成する。(2) 種表現の一部を特定の文字に置換して伏字を生成する。(3) 種表現を別の言語に翻訳した翻訳表現を生成する。

[0063] 同義辞書32は、同義関係にある表現を記憶する辞書であり、同義関係の開始、終了時刻もあわせて登録することができる。同義辞書32に格納された同義関係は、シソーラスを使用した検索、テキストデータのクラス分け、グループ分け、自然語解析、データマイニング、トレンド分析、風評・評判

調査など様々な用途で使うことができる。

[0064] 以下、これを更に具体的に説明する。

図2に示す同義候補リスト10Aは、同義候補EWが多義となる例を含んでいる。同義候補EW[1]に対して、同義元OW[1]から[4]が同義関係の候補として登録されている。同義候補EW[1]は、例えば東西南北の内の一語（例えば「東」）であると、社名、国名など多数の同義元OWが有り得る。

この同義候補リスト10Aには、同義候補EW[1]から[n]まで、同義元OW[1]から[n]までである。同一の種表現[1]である同義元OW[1]が、複数の同義候補EW[1],[2],[3]と関連することもある。

[0065] ここで、同義表現候補記録部10が、複数の同義元OWに対して同義語の候補となる多義の同義候補EWを記憶している際には、同義元検索部18が、複数出現処理18aを備え、同義関係抽出部20が、選択処理20aを備えると良い（図1参照）。

[0066] 図1において、複数出現処理18aは、多義の同義候補EWと同義関係となる同義元OW毎に当該同義元OWの出現を計算する。そして、選択処理20aは、当該多義の同義候補EWの時区間PDと重なる期間の複数の同義元OWの出現を比較することで、当該多義の同義候補EWに対して同義関係となる同義元OWを選択する。

[0067] 例えば、図2に示す同義候補EW[2]の時区間PDに、複数の同義元OW[1],[5]が出現しとする。複数出現処理18aは、時区間PD中、同義元OW[1]の出現と、同義元OW[5]の出現とを算出する。そして、選択処理20aは、同義元OW[1]の出現と、同義元OW[5]の出現とを比較することで、同義関係となる同義元OW[5]を選択する。

[0068] 比較による選択としては、出現数や出現比率が高い同義元OWの選択や、出現数や出現比率が低い同義元OWを候補から除去する選択などがある。選択処理20aは、唯一の同義元OWを選択しても良いし、複数の同義元OWを選択するようにしても良い。

[0069] このように、多義の同義候補EWの時区間PDに複数の同義元OWが検索

された際に、同義関係抽出部 20 が、出現を比較することで同義関係を抽出すると、時区分毎に適した同義元 OW を特定することができる。また、同義関係抽出部 20 は、出現の比較をせず、個々の同義元 OW が一定の条件を満たす同義元 OW との同義関係を成立させるようにしても良い。

[0070] 図 3 に、同義候補 EW の出現と、この同義候補 EW と同義関係を持つ可能性のある同義元 OW [1] から [8] の出現との時間推移を示す。同義元 OW の出現数は図中上方に増加し、同義候補 EW の出現数は図中下方に増加する。時区間 PD は、同義候補 EW が検索されている期間である。図 3 に示す例では、時区間 PD [1] から [6] は同義関係の終了時点を有し、時区間 PD [7] は未終了とした。

[0071] 出現が共通： 時区間 PD [1] から時区間 PD [4]

上記同義関係抽出部 20 は、同義候補 EW と同義元 OW との出現が時間的に共通する際に、同義関係を抽出することができる。ここで、同義候補 EW の時区間 PD に同義元 OW の時区間 PD が重なり合う際には、出現が時間的に共通する。

[0072] 例えば、同義候補 EW の時区間 PD [1] で同義元 OW [1] が出現しているため、同義関係抽出部 20 は、同義候補 EW と同義元 OW [1] とが時区間 PD [1] で成立すると判定する。同様に、同義関係抽出部 20 は、同義候補 EW と、時区間 PD [2] では同義元 OW [2] と、時区間 PD [4] では同義元 OW [3] との同義関係が成立したと判定することができる。

[0073] 時区間 PD [3] については、時区間 PD [3] で出現が共通する同義元 OW を特定できないが、時区間 PD [3] に重なる期間として時区間 PD [3] より前の期間を含めると、同義元 OW [2] との同義関係を抽出することができる。また、同義関係抽出部 20 は、出現の変化率を使用すると、時区間 PD [3] 中に同義元 OW [3] の出現が急増していることから、同義元 OW [3] との同義関係を抽出することもできる。

[0074] (出現が最多の時区間 PD [5] と時区間 PD [6] の場合)

同義関係抽出部 20 は、時区間 PD 中に複数の同義元 OW がある際に、出

現数が最多の同義元OWと同義関係が成立したと判定することができる。

ここで、時区間PD[5]と時区間PD[6]の同義元OWの出現数は、同義元検索部18が、それぞれの時区間PDで特定できるテキスト集合TXから同義元OWを検索し、時区間PDでの出現数を合計した値である。

[0075] そして、同義関係抽出部20は、時区間PD[5]では、棒グラフで示すように出現が最多である同義元OW[5]を選択し、同義候補EWとの同義関係を成立させることができる。また、同一の時区間PDに複数の同義関係を許容する際には、時区間PD[6]に示す例では、予め定められた条件よりも出現数の少ない同義元OW[4]を除去し、同義元OW[5]及び[6]との同義関係を成立させるようにしても良い。

[0076] (出現変化率が高い時区間PD[7]の場合)

同義関係抽出部20は、単位時間当たりの出現数を使用した計算値に基づいて同義関係を抽出することもできる。出現数を単位時間当たりとすることで、同一の出現数の変化率や、異なる同義元OW間の出現数の比などを使用することができる。

例えば、図3に時区分PD[7]との関係で同義元OW[7]と同義元OW[8]の単位時間当たりの出現数の時間変化を示す。図中時区分PD[7]を示す縦線と平行な間隔の短い縦線間の時間が単位時間である。

[0077] 同義元検索部18は、単位時間毎にテキスト集合TXから同義元OWを検索し、その出現数を計算する。時区間PD[7]では、同義元OW[7]の出現数が多く、時区間PD[7]単位の出現数(総数)や、単位時間当たりの出現数を同義元OW[8]と比較すると、同義元OW[7]が選択される。

[0078] 一方、時区間PD[7]と重なる期間で同義元OW[8]の変化率OW[8d]が上昇しており、これに対して、同義元OW[7]の変化率は小さい。同義関係抽出部20は、時区間PD[7]に重なる期間の同義元OWの出現数の変化率に基づいて、最多の同義元OW[7]ではなく、使用が急上昇している同義元OW[8]との同義関係を抽出することができる。

[0079] 同義関係抽出部 20 は、同義候補 EW の時区間 PD と同義元 OW の時区間 PD との共通性（重なり合い等）から、同義候補 EW と同義元 OW との同義関係を判定しても良い。

まず、同義表現候補記録部 10 には、同義元 OW と同義関係となる表現の候補である同義候補 EW が同義元 OW と共に記録されている。

[0080] そして、同義候補検索部 16 は、発行時間の特定可能なテキスト集合 TX を参照して、同義候補 EW がテキスト集合 TX で検出される時区間 PD を算出する。さらに、同義元検索部 18 は、同義元 OW がテキスト集合 TX で検出される時区間 PD を算出する。

[0081] この例では、同義関係抽出部 20 は、同義候補 EW がテキスト集合 TX で検出される時区間 PD と、同義元 OW がテキスト集合 TX で検出される時区間 PD との関係から、同義候補 EW と同義元 OW が同義関係にある時区間 PD を判定する。

[0082] （第 1 実施形態の動作説明）

次に、上記第 1 実施形態における動作を、図 4 に基づいて説明する。

[0083] まず、同義候補生成部 30 は、前述したように種表現から同義候補 EW を生成し同義表現候補記録部 10 に記憶する（図 4：ステップ S101／同義候補生成登録工程）。この場合、同義候補生成部 30 は、ユーザにより作成された同義候補 EW が入力された場合には、これを受け付けて同義表現候補記録部 10 に記憶するように構成してもよい。

[0084] 次に、テキスト収集部 14 は、外部入力されるテキストを収集しこれに基づいて発行時間が特定できるテキスト集合を生成する（図 4：ステップ S102／テキスト集合生成工程）。

そして、この生成された前記テキスト集合に含まれる前記同義表現候補と前記同義元表現との同義関係を、前記同義関係判定特定手段 12 が一定の基準に基づいて判定すると共に特定する（図 4：ステップ S103, S104／同義関係特定工程）。

[0085] ここで、この同義関係を特定する工程（図 4：ステップ S103, S10

4) にあつては、まず、生成された前記ナキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ前記同義関係判定特定手段 12 の同義候補検出部 12A が検索して特定する (図 4: ストップ S 103 / 同義候補検出工程)。

[0086] 続いて、前記同義表現候補が前記ナキスト集合で検出される時区間と前記同義元表現が前記ナキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義関係判定特定手段 12 の同義期間特定部 12B が、前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する (図 4: ストップ S 104 / 同義期間特定工程)。

[0087] ここで、前述した同義候補を検出する工程 (図 4: ストップ S 103) にあつては、まず、同義候補検出部 12A の同義候補検索部 16 が機能して、前記ナキスト収集部 14 で収集された前記ナキスト集合 TX から前記同義表現候補を検出し計数すると共に、単位時間当たりの出現数の大きい時区間 PD を抽出し特定する (図 4: ストップ S 103A / 同義候補対応・時区間特定工程)。

具体的には、同義候補検索部 16 は、ナキスト集合 TX を読んでいき、例えば、同義表現候補記録部 10 に記憶されている同義表現候補 EW の単位期間あたりの出現数が大きく増加している時区間 PD を検出する。

[0088] 又、これと相前後して、同義候補検出部 12A の同義元検索部 18 が機能して、同じく前記ナキスト収集部で収集された前記ナキスト集合から前記同義元表現 OW を検出し、その出現数を計数すると共に単位時間当たりの出現数の大きい時区間が抽出し特定する (図 4: ストップ S 103B / 同義元対応・時区間特定工程)。

[0089] 更に、前述した同義期間を特定する工程 (図 4: ストップ S 104) にあつては、まず、同義期間特定手段 12B の同義関係抽出判定部 20 が機能して、前記同義候補の検出工程 (図 4: ストップ S 103) で前記同義表現候補が検出された時区間におけるナキスト集合中で最も出現数の多い同義元表現を、同義関係にある同義元として抽出する (図 4: ストップ S 104A /

同義関係抽出工程)。

即ち、同義関係抽出部 20 は、同義候補検索部 16 で検出された同義候補 EW がどの同義元 OW と同義関係にあるかを判定し、時区間 PD での同義関係を抽出する。

[0090] 又、これと相前後して、同義区間特定手段 12B の同義期間開始判定部 22 が機能して、同じく前記抽出された同義元表現が前記同義表現候補と同義関係にあると判定し、同時に当該両者が同義関係にある時区間の開始点を同義期間の開始点として、これを前記同義期間と共に予め装備した同義辞書に登録する(図 4 : ステップ S104B / 同義期間登録工程)。

[0091] そして、最後に、同義区間特定手段 12B の同義期間開始判定部 22 が機能して、前述した同義関係ある旨の判定結果を同義辞書 32 に記憶し(図 4 : ステップ S105 / 同義関係登録工程)、これをもって処理を終える。

[0092] このように、本第 1 実施形態では、同義候補 EW が出現した時区間 PD (または時点) での出現(出現数、出現変化率、出現比等)を用いて同義性を判定し、同義関係が成立する時区間 PD を計算する。この同義関係が成立する時区間 PD の特定により、多義で曖昧な語について「時区間同義」を判定し、これにより、時刻によって同義性が変化する場合であっても時区間 PD で区切った同義関係を取り扱うことができる。

[0093] 即ち、本第 1 実施形態では、同義表現候補が出現した時点での単位期間あたりの出現数を用いて同義性の判定を行うため、同義関係が成立した開始時刻を出力することが可能となる。そのため、時刻によって同義性が変化する場合にその同義関係が成立する時区間を判定することが可能となる。

[0094] (同義期間開始について)

次に、同義候補 EW と同義元 OW との同義関係について同義関係の開始時点を判定する手法を開示する。

この例では、図 1 に示すように、同義関係判定特定手段 12 が同義期間開始判定部 22 を備えている。

この同義期間開始判定部 22 は、発行時間順のテキスト集合 TX で同義候

補E Wが検索される時区間P Dにて当該テキスト集合T Xでの同義元O Wの出現が予め定められた条件を満たした時点で当該同義元O Wと同義となる同義期間が開始したと判定する。

[0095] 予め定められた条件は、図3等を参照して開示したように、様々なデータ、比較処理及び判定処理を採用することができる。出現についてのデータとしては、出現数、出現変化率、出現比等を使用することができる。

[0096] 比較及び判定としては、予め定められたしきい値との比較、当該同義元の通常時の出現との比較で求められるしきい値との比較、他の同義元の出現の値との関係で求まる値との比較などがある。

データを比較した結果、条件を満たすと判定しても良いし、実施の具体例によっては例外事項等を定めておき、例えば、しきい値を超えても例外事項に該当する際には条件を満たさないと判定しても良い。

[0097] この同義期間開始判定部22は、例えば、同義候補検索部16で検出された同義候補E Wが、対応する同義元O Wのどれと同義関係にあるかを判定し、同義関係であった場合、同義候補検索部16で検出された時区間P Dの開始点を同義関係の開始点として同義辞書32に登録するようにしても良い。

この例では、同義候補E Wの意味を、開始時点の前後で異なると把握することができる。

[0098] 図1に示す例では、同義関係抽出部20と同義期間開始判定部22とは別の部としているが、同義関係抽出部20が同義期間開始判定部22を備える構成としても良い。

[0099] 同義期間開始判定部22は、出現数で判定するには、まず、同義候補E Wがテキスト集合T Xで検出される時区間P Dにおけるテキスト集合T X中で、最も出現数の多い同義元O Wを同義候補E Wと同義関係にあると判定する。そして、同義期間開始判定部22は、時区間P Dの開始点において同義候補E Wと同義元O Wとの間での同義関係の開始点と判定する。

[0100] 図5に、最多出現数での判定処理の構成例を示す。この例では、同義期間開始判定部22が、この最多出現数での判定をするために、出現数処理22

aと、最多判定処理22bとを備えている。出現数処理22aは、同義候補EWがテキスト集合TXで検索される時区間PDで当該同義候補EWと関連する複数の同義元OWの出現数を計算する。そして、最多判定処理22bが、出現数が最も多い同義元OWと同義候補EWとの同義期間が同義候補EWの時区間PDの開始点で開始したと判定する。

[0101] 図5に示す例では、同義元検索部18が同義元OWの検索結果である出現数を同義元テーブル10Bに記録する。同義期間開始判定部22は、この同義元テーブル10Bを参照して最多判定処理22bをする。

[0102] 図3に示す例では、出現数処理22aは、時区間PD[5]にて複数の同義元OWの出現数を計算する。検索対象の同義元OWは、時区間PD[5]を求めた際同義候補EWに予め記憶手段10にて関連性があるとして記憶されている同義元OW[4],[5],[6]である。図3に示す例では、出現数処理22aは、時区間PD[5]では同義元OW[4],[5],[6]の出現数を計算し、同義元テーブル10Bに記録する。出現数が0の同義元OWについては計算しなくとも良い。

[0103] そして、最多判定処理22bは、図5に示す出現数が90で最多の同義元OW[5]を選択し、同義元OW[5]と同義候補EWとについて、当該時区間PD[5]の開始点で同義関係が開始したと判定する。

この出現数による同義期間開始判定は、通常時から注目されている同義元OWとの同義関係の判定に効果的である。

[0104] 次に、出現率を使用する処理例を説明する。

同義期間開始判定部22は、出現比率で判定するには、時区間PD中での単位時間あたりの出現数と、時区間PD以前での単位時間あたりの出現数との出現比率が最も大きい同義元OWを同義候補EWと同義関係にあると判定する。そして、同義期間開始判定部22は、時区間PDの開始点において同義候補EWと同義元OWとの間での同義関係の開始点と判定する。

[0105] 図6に出現比率での判定処理の一例を示す。この例では、同義期間開始判定部22が、この出現比率での判定をするために、時区間中処理22cと、

時区間前処理 2 2 d と、比率判定処理 2 2 e とを備えている。

時区間中処理 2 2 c は、同義候補 E W がテキスト集合 T X で検索される時区間 P D での当該同義候補 E W と関連する 1 以上の同義元 O W の単位時間あたりの出現数を計算する。時区間前処理 2 2 d は、当該各同義元 O W の時区間 P D 前での単位時間あたりの出現数を計算する。

[0106] そして、比率判定処理 2 2 e は、時区間 P D 前の出現数に対する時区間 P D 中の出現数の比率が最も大きい同義元 O W との同義期間が、同義候補 E W の時区間 P D の開始点で開始したと判定する。

即ち、比率判定処理 2 2 e は、時区間 P D の開始時点より前の出現数と比較して、時区間 P D 中の出現数が大きい際に、同義期間が同義候補 E W の時区間 P D の開始点で開始したと判定する。

[0107] 図 6 に示す例では、同義元検索部 1 8 が、同義候補 E W の時区間 P D 前の出現数も計算し、同義元 O W の出現数を同義元テーブル 1 0 B に格納する。さらに、比率判定処理 2 2 e が、算出した出現比率を同義候補テーブル 1 0 C に格納する。

[0108] 図 7 及び図 8 のフローチャートを参照し、図 3 に示す時区間 P D [7] を例として出現比率による処理工程を説明する。

まず、テキスト収集部 1 4 は、予め定められた周期や時刻（検索時刻）にネットワーク 9 6 を介してテキストを検索し、テキスト集合 T X を生成する（図 7：ステップ S 2 0 1）。テキスト収集部 1 4 は、さらに、テキストの発行時間を特定する（図 7：ステップ S 2 0 2）。

[0109] 続いて、同義元検索部 1 8 は、記憶手段 1 0 に登録されている全ての同義元 O W について順次検索する（図 7：ステップ S 2 0 3, S 2 0 7）。同義元 O W が検索されると、同義元検索部 1 8 は、テキスト集合 T X 中の単位時間あたりの出現数を算出し（図 7：ステップ S 2 0 4）、同義元テーブル 1 0 B に記録する（図 7：ステップ S 2 0 5）。

[0110] 収集したテキスト集合 T X に対して全ての同義元 O W の検索が終了すると、次のテキスト収集を待機し、出現数を算出していない同義元があれば、次

の同義元を特定して (図 7 : ストップ S 2 0 7) 、 検索する (図 7 : ストップ S 2 0 3) 。

同義元検索部 1 8 は、同義候補 E W と時区間 P D [7] とが特定されている際には、図 6 の同義元テーブル 1 0 B に示すように、時区間 P D [7] の前の区間の同義元 O W [8] の出現数 (1 0 0) と、時区間 P D [7] 中の同義元 O W [8] の出現数 (4 0 0) とを記録する。

[0111] 図 8 のフローチャートにあって、同義候補検索部 1 6 は、まず、同義表現候補記録部 1 0 に登録されている同義候補 E W を順次検索する (図 8 : ストップ S 2 1 1, S 2 1 3) 。同義候補 E W が予め定められた数以上発見されたと (図 8 : ストップ S 2 1 2) 、同義期間開始判定部 2 2 は、発見したテキストの発行時間を開始時点として、同義候補 E W の時区間 P D を開始する。図 3 に示す例では、時区間 P D [7] を開始する。

[0112] 続いて、同義期間開始判定部 2 2 の時区間中処理 2 2 c は、同義候補 E W が検索された時区間 P D [7] での同義元 O W の出現数を計算し (図 8 : ストップ S 2 1 6) 、同義元テーブル 1 0 B に記録する。ここでは、図 3 に示す最短の単位時間ではなく、一定時間まとめた総数 (4 0 0) とすると良い。

そして、時区間前処理 2 2 d は、当該各同義元 O W の時区間 P D 前での同様の一定時間まとめた出現数 (1 0 0) を計算し (図 8 : ストップ S 2 1 7) 、同義元テーブル 1 0 B に記録する。

[0113] 更に、比率判定処理 2 2 e は、時区間 P D 前の出現数 (1 0 0) に対する時区間 P D 中の出現数 (4 0 0) の比率 (4 0 0 %) を算出する (ストップ S 2 1 8) 。この比率判定処理 2 2 e は、さらに、出現比率が最も大きい同義元 O W [8] を選択し、この同義元 O W [8] との同義期間が同義候補 E W の時区間 P D [7] の開始点で開始したと判定する (図 8 : ストップ S 2 2 1) 。

[0114] そして、同義期間開始判定部 2 2 は、この時区分同義である同義関係を同義辞書 3 2 に記録する (図 8 : ストップ S 2 2 2) 。

[0115] 図 8 に示す出現比率による同義期間の開始判定処理は、通常時には注目度が低い同義元 O W との同義関係の抽出に効果的である。

[0116] ここで、上述した各構成の動作部分における動作内容（特に図４，図７及び図８における各動作ステップ）については、これをコンピュータで実行可能にプログラム化し、これらを前述した各ステップを実行する同義関係判定特定手段１２が備えているコンピュータに実行させるようにしてもよい。他の実施形態においても同様である。

[0117] この場合、プログラム化したプログラムについては、非一時的な記録媒体、例えば、ＤＶＤやＣＤ、フラッシュメモリなどに記録させてもよい。その場合、本プログラムは、記録媒体からコンピュータによって読み出され、実行される。

[0118] 上述したように、本第１実施形態によると、前述したように同義元ＯＷの出現を使用した情報処理により、時区分同義を判定し、特に、開始時間が明確な時区分同義を判定することができる。

〔第２実施形態〕

[0119] 次に、本発明の第２実施形態を図９乃至図１０に基づいて説明する。

まず、図９に示す第２実施形態にあって、同義関係判定装置１０２は、前述した図１に開示した第１実施形態の各構成に加えて同義期間の終了を判定する同義期間終了判定部２４を備えている点に特徴を有する。

[0120] 即ち、この第２実施形態では、同義関係判定特定手段１２が、発行時間順のテキスト集合ＴＸにて同義候補ＥＷの出現が予め定められた条件未滿に減少した時点で同義期間が終了したと判定する同義期間終了判定部２４を備えている。そして、この同義期間終了判定部２４が、同義候補ＥＷの出現が減少した時点で同義関係が終了したと判定するため、多義的で曖昧な同義候補ＥＷの意味が成立する期間を一定期間内に特定することができる。

[0121] 図１０は、本第２実施形態における同義関係判定装置１０２による情報処理の一例を示すフローチャートである。

まず、同義表現生成部３０で、種表現から同義候補ＥＷを生成し、同義表現候補記録部１０に記録する（図１０：ステップＳ２２１）。

[0122] 次に、同義候補検索部１６で、テキスト集合ＴＸを読んでいき、同義表現

候補記録部 10 に記録されている同義候補 EW の単位期間あたりの出現数が大きく増加している時区間 PD を検出する（図 10：ステップ S 222）。

そして、同義期間開始判定部 22 で、同義候補検索部 16 で検出された同義候補 EW がどの同義元 OW と同義関係にあるかを判定し、同義関係の開始点を判定する（図 10：ステップ S 223）。

[0123] この図 10 に示す例では、同義期間終了判定部 24 が、同義期間開始判定部 22 で同義関係にあると判定された同義候補 EW の単位期間あたりの出現数が終了しきい値以下になった場合に、同義関係が解消していると判定し（図 10：ステップ S 224）、同義辞書 32 に当該同義関係の終了時刻を登録する（図 10：ステップ S 225）。

[0124] 次に、本第 2 実施形態における効果について説明する。

本実施の形態では、同義関係が終了時刻を出力することが可能であるため、同義関係が成立している時区間 PD を正しく判定することが可能となる。

〔第 3 実施形態〕

[0125] 次に、本発明の第 3 実施形態を図 11 乃至図 12 に基づいて説明する。

この第 3 実施形態では、時区間 PD 内の同義元を 1 つとし、同義関係の開始、入替及び終了を判定する点に特徴を有する。

[0126] まず、図 11 に示す第 3 実施形態にあって、同義関係判定装置 103 は、同義関係抽出部 20 が、出現計算処理 20b と、開始判定処理 20c と、入替処理 20d と、終了判定処理 20e とを備えている。

[0127] 出現計算処理 20b は、同義候補 EW がテキスト集合 TX で検索された際に当該同義候補 EW と関連する複数の同義元 OW の出現を計算する。

開始判定処理 20c は、出現が予め定められた開始しきい値を超えた同義元 OW のうち当該出現が最多の同義元 OW について、当該開始しきい値を超えた時点で同義候補 EW との同義関係が開始したと判定する。

[0128] 入替処理 20d は、同義関係開始後に当該同義元 OW の出現が他の同義元 OW の出現を下回った際に、当該下回った時点で当該同義関係が終了したと判定すると共に新たに最多の同義元 OW について同義関係が開始したと判定

する。

- [0129] 終了判定処理 20e は、同義関係開始後に出現が予め定められた終了しきい値を下回った際に、当該下回った時点で当該同義関係が終了したと判定する。

この構成により、多義的な同義候補 EW の意味をより高確率な意味に特定することができる。

その他の構成は、前述した第 1 実施形態の場合と同一となっている。

- [0130] これを、更に詳細に説明する。

ここで、いま、種表現（同義元）として、図 12 に示すように「東京電力」と「東北電力」が与えられた場合を想定する。

- [0131] まず、同義候補生成部 30 で、同義元 OW の形態素の先頭文字を残して省略語を生成したり、同義元 OW の 1 文字を「○」に置換して伏字を生成したりすることで同義候補 EW を複数生成する。たとえば、種表現（同義元 OW）が「東京電力」の場合、「東電」や「東京電」のような省略語、「東○電力」のような伏字が生成される。

- [0132] このように「東京電力」（同義元 OW[10]）の伏字として「東○電力」（同義候補 EW）が考えられるが、これは、「東北電力」（同義元 OW[11]）の伏字にもなりうる。このように、「東○電力」（同義候補 EW）が指す内容は「東京電力」（同義元 OW[10]）や「東北電力」（同義元 OW[11]）等の曖昧性がある。

- [0133] また、種表現（同義元 OW）が「東北電力」の場合は、「東電」や「東北電」のような省略語、「東○電力」のような伏字が生成される。以上のようにして、図 12 に示す同義候補リスト 10A 又はこのようなデータが記憶手段 10 に記憶される。

- [0134] このとき、「東電」、「東○電力」は、「東京電力」、「東北電力」のいずれからも生成されており、上述のように曖昧性がある同義候補 EW となっている。

そして、実際には、時刻によって「東○電力」が指す内容が「東京電力」

や「東北電力」に変化しうる。

[0135] 例えば、図 1 3 に示す例では、「東○電力」（同義候補 E W）は、時区間 P D[A]、時区間 P D[C]では「東京電力」（同義元 O W[10]）を、時区間 P D[B]では「東北電力」（同義元 O W[11]）を指しており、同義関係が時刻によって変化している。

[0136] ここで、本第 3 実施形態の情報処理例を前述した図 1 1 を参照して説明する。

まず、同義候補生成部 3 0 で、種表現中の形態素の先頭文字を残して省略語を生成したり、種表現中の 1 文字を「○」に置換して伏字を生成したりすることで同義候補 E W を複数生成する。

次に、同義候補検索部 1 6 及び同義元検索部 1 8 で、同義候補 E W 及び同義元 O W がテキスト集合 T X 中で（突発的に）出現する時点を検出する。テキスト集合 T X 中の各テキストには、クロール時間、書き込み時間、等の発行時間が付与されており、それをもとに同義候補 E W 及び同義元 O W が出現する時点を検出する。

[0137] そして同義関係抽出部 2 0 は、テキスト集合 T X 中で「東京電力」（同義元 O W[11]）、「東北電力」（同義元 O W[11]）、「東○電力」（同義候補 E W）の出現の頻度を計数し（出現計算処理 2 0 b）、図 1 3 に示すような出現頻度になっていた場合、図 1 3 の時区間 P D[A]、時区間 P D[B]、及び時区間 P D[C]を算出する。

[0138] 同義関係抽出部 2 0 は、同義性を判定する。例えば、図 1 3 の時区間 P D[A]では、「東○電力」の同義元 O W は、「東京電力」と「東北電力」の 2 つがあるが、時区間 P D[A]での出現数が「東京電力」では、1 日あたり 8 0 0 個、「東北電力」では、1 日あたり 1 5 0 個だとすると、「東京電力」のほうが多く出現するため、時区間 P D[A]の開始時点から「東○電力」と「東京電力」が同義関係であると判定し（開始判定処理 2 0 c）、同義辞書 3 2 に時区間 P D[A]の開始時点から「東○電力」と「東京電力」が同義関係にあると登録される。

[0139] 同様に、同義関係抽出部 20 は、時区間 P D [B] では、「東○電力」と「東

北電力」が（入替処理 20 d）、時区間 P D [C] では、「東○電力」と「東京

電力」が同義関係であると判定する（入替処理 20 d）。

[0140] 以上のように「東○電力」は時期によって同義関係が変化するが、それに

応じて同義辞書 32 も更新されることになり、時期によって意味が変化する

同義関係の成立する時区間 P D を正しく判定することが可能となる。

[0141] 続いて、同義関係抽出部 20 は、同義関係が成立した後の出現数を監視し

、図 13 の時区間 P D [D] のように、「東○電力」の出現がしきい値以下に減

少し、「東○電力」と同義関係にある「東京電力」の単位期間あたりの出現

数も通常と同レベルにまで減少した場合に、「東○電力」と「東京電力」の

同義関係が終了したと判断し（終了判定処理 20 e）、同義辞書 32 の当該

項目に終了時刻とともに同義関係が終了したことを登録する。

これにより、同義候補 E W の出現が少なくなり、同義関係が成立すると

えなくなつた場合に、終了時刻とともに同義関係が終了したことを判定する

ことができる。

[0142] 図 14 は、第 3 実施形態での別の情報処理例を示すフローチャートである

。

この図 14 に示す処理では、上述の情報処理とはしきい値や終了の扱い等

が異なるが情報処理の概要は同様である。

ここでは、同義元 O W の単位時間毎の出現については、図 7 に示すフロ

ーチャートで計算され、同義元テーパー 10 B 等に格納されているとする。

[0143] まず、図 8 に示す処理と同様に、同義関係抽出部 20 が、記憶手段 10 に

記憶されている同義候補 E W を検索し（図 14：ステップ S 211）、同義

候補 E W を発見しない際には（図 14：ステップ S 212）、次の同義候補

E W を特定して（図 14：ステップ S 213）、再度検索する（図 14：ス

テップ S 211）。

[0144] 図 13 に示す「東○電力」が検索されたとすると（図 14：ステップ S 2

12）、同義関係抽出部 20 は、図 12 に示す同義候補リスト 10 A を参照

して「東○電力」に対応する同義元OWである「東京電力」（同義元OW[10]）と「東北電力」（同義元OW[11]）とを特定する（図14：ステップS215）。

[0145] 上記出現計算処理20bは、「東○電力」（同義候補EW）がテキスト集合TXで検索された（ステップS212）後、関連する複数の同義元OW[10],[11]のそれぞれの出現を計算する（図14：ステップS301）。

[0146] そして、開始判定処理20cは、複数の同義元OW[10],[11]のうち出現が予め定められた開始しきい値を超えたか否かを確認する（図14：ステップS302）。開始判定処理20cは、続いて、開始しきい値を超えている際には、同義元OW群のうち当該出現が最多の「東京電力」（同義元OW[10]）を選択する（図14：ステップS303）。そして、開始判定処理20cは、当該開始しきい値を超えた時点で同義候補EWとの同義関係が開始したと判定する（図14：ステップS304）。

[0147] 図13に示す例では、時区間PD[B]に至ると、同義元OW[10]の出現数が減少して、同義元OW[11]の出現数が増加している。

入替処理20dは、同義関係開始後に、当該同義元OW[10]（東京電力）の出現（現出現）が他の同義元OW[11]（東北電力）の出現（他出現）を下回った際に（図14：ステップS305）、当該下回った時点で当該同義関係が終了したと判定し、そして、新たに最多の同義元OWについて同義関係が開始したと判定する（図14：ステップS306）。現出現が他出現より多い状態が継続すれば（図14：ステップS305）、この入替処理20dは実行されず、終了処理の判定に移る。

[0148] この図13に示す例では、「東北電力」との入替による同義関係の開始後、出現数が終了しきい値を下回らないまま（図14：ステップS307）、時区間PD[C]に至ったとする。ここでは、再度、他出現である「東京電力」との同義関係の入替がなされる（図14：ステップS305、S306）。

[0149] 東京電力との同義関係が継続したまま、時区間PD[D]に至ると、同義元OW[10]の出現が終了しきい値を下回る（図14：ステップS307）。この

場合、終了判定処理 20e が、当該下回った時点で当該同義関係が終了したと判定する（図 14：ステップ S308）。

[0150] 次に、同義関係を同義辞書 32 に格納するのみならず、表示デバイスに表示制御する例を説明する。

前述した図 11 において、実施例 3 の同義関係判定装置 103 は、同義関係判定特定手段 12 に、データを表示する表示デバイス 95 を併設している。そして、同義関係判定特定手段 12 が、表示制御部 26 を備えている。

[0151] この表示制御部 26 は、同義関係抽出部 20 によって抽出された同義候補 EW と、同義関係開始時点と、当該同義関係開始時点で同義関係が開始した同義元 OW と、同義関係終了時点とを時区分別同義データ TD として表示制御する。これにより、同義関係の時間変化の情報を含めてテキスト集合 TX 等をユーザに表示することができる。

[0152] さらに、時区間同義によりフィルタリングしたテキスト集合 TX を表示すると、ユーザが手作業で個々に同義候補 EW との対応等を検討して検索することと比較して、大幅に作業の生産性を向上させることができる。

[0153] 本発明の活用例として、インターネットを対象とした風評監視システム、評判抽出システム、等に適用できる。

[0154] ＜実施形態の全体的な動作＞

ここで、上記実施例 1 乃至実施例 3 の同義関係判定装置 101, 102, 103 に共通する情報処理について、ハードウェア資源を参照して説明する。

[0155] （ハードウェア資源）

本実施形態における同義関係判定装置 101, 102, 103 による情報処理は、ソフトウェアとハードウェア資源とが協働し、使用目的に応じて情報を演算し、又は加工する具体的手段である。

ハードウェア資源として、図 15 に示すように、情報処理をするコンピュータ 80 を有している。コンピュータ 80 は、中央処理装置（CPU）である演算手段 82 と、この演算手段 82 に記憶領域を提供する主記憶手段 86

を有する。コンピュータ 80 は、一般に、データバス及び入出力インタフェースを通じて接続される周辺機器を有する。周辺機器は、代表的には、通信手段 88、外部記憶手段 90、入力手段 92、出力手段 94 である。周辺機器を含めた全体をコンピュータ 80 ということもある。

[0156] 通信手段 88 は、有線又は無線のネットワークを介してサーバ装置 70 との通信を制御する。外部記憶手段 90 は、プログラムファイル 100 や、データを記憶する据え付け又は持ち運び可能な記憶媒体である。入力手段 92 はキーボード、タッチパネル、ポインティングデバイス、スキャナ等でありユーザの操作に応じてコンピュータ 80 で読み取り可能なデータを入力する。出力手段 94 は、ディスプレイ、プリンタなどコンピュータ 80 が計算したデータ等を表示出力する。

[0157] 本実施形態による同義関係判定装置 101、102、103 の記憶手段 10 は、外部記憶手段 90 をハードウェア資源として使用して同義候補リスト 10A 等のデータを記憶する。また、同義関係判定特定手段 12 は、CPU である演算手段 82 をハードウェア資源として使用して、テキスト集合 TX に対するデータ処理を実行する。すなわち、同義関係判定特定手段 12 は、プログラムを実行するコンピュータ 80 で実現することができる。

[0158] 上述した各実施形態について、その新規な技術内容の要点をまとめると、以下ようになる。尚、上記実施形態の一部又は全部は、新規な技術としては以下のようにまとめられるが、本発明は必ずしもこれに限定されるものではない。

[0159] [付記 1]

所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部 10 と、外部入力されるテキスト中における前記同義表現候補と前記同義元表現との同義関係を一定の基準に基づいて判定し特定する同義関係判定特定手段 12 とを備えた同義関係判定装置であって、

前記同義関係判定特定手段 12 が、

外部入力される前記テキストを収集しこれに基づいて発行時間が特定可能なテキスト集合を生成するテキスト収集部 14 と、

このテキスト収集部 14 で収集されたテキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを特定し出力する同義候補検出部 12 A と、

前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する同義期間特定部 12 B と、

を備えたことを特徴とする同義関係判定装置。

[0160] [付記 2]

付記 1 に記載の同義関係判定装置において、

前記同義候補検出部 12 A を、前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を（前記同義表現候補が存在する時区間として）特定する同義候補検索部 16 と、同じく前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を（前記同義元表現が存在する時区間として）特定する同義元検索部 18 とを含む構成としたことを特徴とする同義関係判定装置。

[0161] [付記 3]

付記 1 又は 2 に記載の同義関係判定装置において、

前記同義期間特定部 12 B を、前記同義候補検出手段で前記同義表現候補が検出された時区間におけるテキスト集合中で少なくとも最も出現数の多い同義元表現を同義関係にある同義元として抽出する同義関係抽出部 20 と、この抽出された同義元表現が前記同義表現候補と同義関係にあると判定すると共に当該両者が同義関係にある時区間の開始点を同義期間の開始点としてこれを同義期間と共に予め装備した同義辞書に登録する同義期間開始判定部

２２とを含む構成としたことを特徴とする同義関係判定装置。

[0162] [付記４]

付記３に記載の同義関係判定装置において、

前記同義期間開始判定部２２は、前記同義表現候補が前記テキスト集合で検出される時区間における集合中で、前記時区間での単位時間あたりの出現数と前記時区間以前での単位時間あたりの出現数の比率が最も大きい同義表現について、これを前記同義表現候補と同義関係にあると判定する機能を備えていることを特徴とした同義関係判定装置。

[0163] [付記５]

付記１，２，３，又は４に記載の同義関係判定装置において、

前記同義期間特定部１２Ｂは、

当該同義期間特定部１２Ｂの前記同義期間開始判定部２２で同義関係にあると判定され特定された同義表現候補についてその単位期間あたりの出現数が予め設定したしきい値以下になった時点で、前記同義関係が解消した旨判定する同義期間終了判定部２４を備えていることを特徴とした同義関係判定装置。

[0164] [付記６] (方法の発明／付記１対応)

所定の１の同義表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部１０と、外部入力されるテキスト中における前記同義表現候補と同義表現とを判定し特定する同義関係判定特定手段１２を備えた同義関係判定装置において、外部入力される前記テキストを収集しこれに基づいて発行時間が特定できるテキスト集合を、前記同義関係判定特定手段１２のテキスト収集部が生成し(テキスト集合生成工程)、

この生成された前記テキスト集合に含まれる前記同義表現候補と前記同義表現との同義関係を、前記同義関係判定特定手段１２が一定の基準に基づいて判定すると共に特定し(同義関係特定工程)、

前記同義関係を特定する工程にあっては、

前記テキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ前記同義関係判定特定手段 1 2 の同義候補検出部 1 2 A が検索して特定し（同義候補検出工程）、

続いて、前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義関係判定特定手段 1 2 の同義期間特定部 1 2 B が前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定するようにしたこと（同義期間特定工程）を特徴とする同義関係判定方法。

[0165] [付記 7]（付記 2 対応）

付記 6 に記載の同義関係判定方法において、

前記同義候補を検出する工程にあつては、

前記テキスト収集部で収集された前記テキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定し（同義候補対応・時区間特定工程）、

これと相前後して同じく前記テキスト収集部で収集された前記テキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定する構成（同義元対応・時区間特定工程）とし、

これらの各時区間の特定工程における動作内容を前記同義候補検出部 1 2 A が実行することを特徴とした同義関係判定方法。

[0166] [付記 8]（付記 3 対応）

付記 6 に記載の同義関係判定方法において、

前記同義期間を特定する工程にあつては、

前記同義候補の検出工程で前記同義表現候補が検出された時区間におけるテキスト集合中で少なくとも最も出現数の多い同義元表現を同義関係にある同義元として抽出し（同義関係抽出工程）、

これと相前後して、同じく前記抽出された同義元表現が前記同義表現候補と同義関係にあると判定し且つ当該両者が同義関係にある時区間の開始点を

同義期間の開始点としてこれを前記同義期間と共に予め装備した同義辞書に登録する構成とし（同義期間登録工程）、

これらの各抽出／登録工程の動作内容を前記同義区間特定手段 1 2 B が実行することを特徴とした同義関係判定方法。

[0167] 〔付記 9〕（付記 5 対応）

付記 6, 7 又は 8 に記載の同義関係判定方法において、

前記同義期間特定部 1 2 B で同義関係にあると判定された同義表現候補の単位期間あたりの出現数を継続的に計数すると共にその出現数が予め設定したしきい値以下になった場合には、前記同義期間特定部 1 2 B の同義期間終了判定部 2 4 が前記同義関係が解消したと判定する構成としたことを特徴とする同義関係判定方法。

[0168] 〔付記 10〕（プログラム発明／付記 6 対応）

所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部 10 と、外部入力されるテキスト中における前記同義表現候補と同義元表現との同義関係を判定し特定する同義関係判定特定手段 1 2 とを備えた同義関係判定装置にあって、

外部入力されるテキストを収集して発行時間を特定できるテキスト集合を生成するテキスト集合生成処理機能、

および生成された前記テキスト集合に含まれる前記同義表現候補と前記同義元表現との同義関係を、一定の基準に基づいて判定すると共に特定処理する同義関係特定処理機能を設けると共に、

前記同義関係特定処理機能が、

前記テキスト収集部で収集された前記テキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ検索して特定処理する同義候補検出処理機能、

および前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義表現候補と前記同義元表現とが同義関係にある時区

間を同義期間と判定し特定する同義期間特定処理機能を含む構成とし、

これらの各処理機能を前記同義関係判定特定手段 1 2 が備えているコンピュータに実現させるようにしたことを特徴とした同義関係判定プログラム。

[0169] [付記 1 1] (付記 7 対応)

付記 1 0 に記載の同義関係判定プログラムにおいて、

前記同義候補検出処理機能が、

前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定する同義候補対応・時区間特定処理機能、

および前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定する同義元対応・時区間特定処理機能、を備えた構成とし、

これら各時区間特定処理機能を前記同義関係判定特定手段 1 2 が備えているコンピュータに実現させるようにしたことを特徴とする同義関係判定プログラム。

[0170] [付記 1 2] (付記 8 対応)

付記 1 0 に記載の同義関係判定プログラムにおいて、

前記同義期間特定処理機能が、

前記同義候補検出処理機能で前記同義表現候補が検出された時区間におけるテキスト集合中で少なくとも最も出現数の多い同義元表現を同義関係にある同義元として抽出する同義関係抽出処理機能、

およびこれと相前後して、同じく前記抽出された同義元表現が前記同義表現候補と同義関係にあると判定し且つ当該両者が同義関係にある時区間の開始点を同義期間の開始点としてこれを前記同義期間と共に予め装備した同義辞書に登録する同義期間登録処理機能、を備えた構成とし、

これらの各処理機能を前記同義関係判定特定手段 1 2 が備えているコンピュータに実現させるようにしたことを特徴とする同義関係判定プログラム。

[0171] [付記 1 3] (付記 9 対応)

付記 1 0, 1 1 又は 1 2 に記載の同義関係判定プログラムにおいて、
前記同義関係特定処理機能が、

前記同義関係にあると判定された同義表現候補の単位期間あたりの出現数を継続的に計数すると共にその出現数が予め設定したしきい値以下になった場合に、その時点で前記同義関係が解消したと判定する同義期間終了判定機能を備え、

これを前記同義関係判定特定手段 1 2 が備えているコンピュータに実現させるようにしたことを特徴とする同義関係判定プログラム。

[0172] この出願は 2 0 1 2 年 3 月 3 0 日に出願された日本出願特願 2 0 1 2 - 0 8 2 7 2 2 を基礎とする優先権を主張し、その開示の全てをここに取り込む。

産業上の利用可能性

[0173] 本発明は、同義関係を利用する全ての自然語データ処理に対して適用可能である。

符号の説明

- [0174] 1 0 記憶手段
1 2 同義関係判定特定手段
1 2 A 同義候補検出部
1 2 B 同義期間特定部
1 4 テキスト収集部
1 6 同義候補検索部
1 8 同義元検索部
1 8 a 複数出現処理
2 0 同義関係抽出部
2 2 同義期間開始判定部
2 4 同義期間終了判定部
2 6 表示制御部

3 0 同義候補生成部

3 2 同義辞書

E W 同義候補

O W 同義元

P D 時区間

D 時区分別同義データ

T X テキスト集合

請求の範囲

- [請求項1] 所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部と、外部入力されるテキスト中における前記同義表現候補と前記同義元表現との同義関係を一定の基準に基づいて判定し特定する同義関係判定特定手段とを備えた同義関係判定装置であって、
- 前記同義関係判定特定手段が、
- 外部入力される前記テキストを収集しこれに基づいて発行時間が特定可能なテキスト集合を生成するテキスト収集部と、
- このテキスト収集部で収集されたテキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを特定し出力する同義候補検出部と、
- 前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する同義期間特定部と、
- を備えたことを特徴とする同義関係判定装置。
- [請求項2] 請求項1に記載の同義関係判定装置において、
- 前記同義候補検出部を、前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を前記同義表現候補が存在する時区間として特定する同義候補検索部と、同じく前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を前記同義元表現が存在する時区間として特定する同義元検索部とを含む構成としたことを特徴とする同義関係判定装置。
- [請求項3] 請求項1又は2に記載の同義関係判定装置において、
- 前記同義期間特定部を、前記同義候補検出部で前記同義表現候補が

検出された時区間におけるテキスト集合中で少なくとも最も出現数の多い同義元表現を同義関係にある同義元として抽出する同義関係抽出部と、この抽出された同義元表現が前記同義表現候補と同義関係にあると判定すると共に当該両者が同義関係にある時区間の開始点を同義期間の開始点としてこれを同義期間と共に予め装備した同義辞書に登録する同義期間開始判定部とを含む構成としたことを特徴とする同義関係判定装置。

[請求項4] 請求項3に記載の同義関係判定装置において、

前記同義期間開始判定部は、前記同義表現候補が前記テキスト集合で検出される時区間における集合中で、前記時区間での単位時間あたりの出現数と前記時区間以前での単位時間あたりの出現数の比率が最も大きい同義元表現についても、これを前記同義表現候補と同義関係にあると判定する機能を備えていることを特徴とした同義関係判定装置。

[請求項5] 請求項1, 2, 3, 又は4に記載の同義関係判定装置において、

前記同義期間特定部は、

当該同義期間特定部の前記同義期間開始判定部で同義関係にあると判定され特定された同義表現候補についてその単位期間あたりの出現数が予め設定したしきい値以下になった時点で、前記同義関係が解消した旨判定する同義期間終了判定部を備えていることを特徴とした同義関係判定装置。

[請求項6] 所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部と、外部入力されるテキスト中における前記同義表現候補と同義元表現との同義関係を判定し特定する同義関係判定特定手段を備えた同義関係判定装置であって、

外部入力される前記テキストを収集しこれに基づいて発行時間が特定できるテキスト集合を、前記同義関係判定特定手段のテキスト収集

部が生成し、

この生成された前記テキスト集合に含まれる前記同義表現候補と前記同義元表現との同義関係を、前記同義関係判定特定手段が一定の基準に基づいて判定すると共に特定し、

前記同義関係を特定する工程にあつては、

前記テキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ前記同義関係判定特定手段の同義候補検出部が検索して特定し、

次に、前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義関係判定特定手段の同義期間特定部が前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定するようにしたことを特徴とする同義関係判定方法。

[請求項7]

請求項6に記載の同義関係判定方法において、

前記同義候補を検出する工程にあつては、

前記テキスト収集部で収集された前記テキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定し、

これと相前後して同じく前記テキスト収集部で収集された前記テキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定する構成とし、

これらの各時区間の特定工程における動作内容を前記同義候補検出部が実行することを特徴とした同義関係判定方法。

[請求項8]

請求項6に記載の同義関係判定方法において、

前記同義期間を特定する工程にあつては、

前記同義候補の検出工程で前記同義表現候補が検出された時区間におけるテキスト集合中で少なくとも最も出現数の多い同義元表現を同

義関係にある同義元として抽出し、

これと相前後して、同じく前記抽出された同義元表現が前記同義表現候補と同義関係にあると判定し且つ当該両者が同義関係にある時区間の開始点を同義期間の開始点としてこれを前記同義期間と共に予め装備した同義辞書に登録する構成とし、

これらの各抽出／登録工程の動作内容を前記同義期間特定手段が実行することを特徴とした同義関係判定方法。

[請求項9]

所定の一の同義元表現と同義関係の対象となる複数の同義表現候補とが相互に対応して記録された同義表現候補記録部と、外部入力されるテキスト中における前記同義表現候補と同義元表現との同義関係を判定し特定する同義関係判定特定手段とを備えた同義関係判定装置にあって、

外部入力されるテキストを収集して発行時間を特定できるテキスト集合を生成するテキスト集合生成処理機能、

および生成された前記テキスト集合に含まれる前記同義表現候補と前記同義元表現との同義関係を、一定の基準に基づいて判定すると共に特定処理する同義関係特定処理機能を設けると共に、

前記同義関係特定処理機能が、

前記テキスト収集部で収集された前記テキスト集合から前記同義表現候補が多く検出される時区間と前記同義元表現が多く検出される時区間とを、それぞれ検索して特定処理する同義候補検出処理機能、

および前記同義表現候補が前記テキスト集合で検出される時区間と前記同義元表現が前記テキスト集合で検出される時区間との位置関係及び検出頻度とに基づいて、前記同義表現候補と前記同義元表現とが同義関係にある時区間を同義期間と判定し特定する同義期間特定処理機能を含む構成とし、

これらの各処理機能を前記同義関係判定特定手段が備えているコンピュータに実現させるようにしたことを特徴とした同義関係判定プロ

グラム。

[請求項10]

請求項9に記載の同義関係判定プログラムにおいて、

前記同義候補検出処理機能が、

前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義表現候補を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定する同義候補対応・時区間特定処理機能、

および前記テキスト収集部で収集された発行時間を特定できるテキスト集合から前記同義元表現を検出し計数すると共に単位時間当たりの出現数の大きい時区間を抽出して特定する同義元対応・時区間特定処理機能、を備えた構成とし、

これら各時区間特定処理機能を前記同義関係判定特定手段が備えているコンピュータに実現させるようにしたことを特徴とする同義関係判定プログラム。