(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
2014 年 1 月 3 日(03.01.2014)

WIPO | PCT

(10) 国際公開番号
WO 2014/002776 A1

(51) 国際特許分類:
G06F 17/30 (2006.01) *G06F 17/27* (2006.01)

(21) 国際出願番号: PCT/JP2013/066286

(22) 国際出願日: 2013 年 6 月 6 日(06.06.2013)

(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 2012-141680 2012 年 6 月 25 日(25.06.2012) JP

(71) 出願人: 日本電気株式会社(NEC CORPORATION) [JP/JP]; 〒1088001 東京都港区芝五丁目 7 番 1 号 Tokyo (JP). 国立大学法人名古屋大学(NATIONAL UNIVERSITY CORPORATION NAGOYA UNIVERSITY) [JP/JP]; 〒4648601 愛知県名古屋市千種区不老町 1 番 Aichi (JP).

(72) 発明者: 平尾 英司(HIRAO, Eiji); 〒1088001 東京都港区芝五丁目 7 番 1 号日本電気株式会社内 Tokyo (JP). 古橋 武(FURUHASHI, Takeshi); 〒4648601 愛知県名古屋市千種区不老町 1 番国立大学法人名古屋大学内 Aichi (JP). 吉川 大弘(YOSHIKAWA, Tomohiro); 〒4648601 愛知県名古屋市千種区不老町 1 番国立大学法人名古屋大学内 Aichi (JP).

(74) 代理人: 池田 憲保, 外(IKEDA, Noriyasu et al.); 〒1000011 東京都千代田区内幸町 1 丁目 2 番 2 号 日比谷ダイビル Tokyo (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, BN, BR, BW, BY, BZ, CA, CH, CL, 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, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[続葉有]

(54) **Title**: SYNONYM EXTRACTION SYSTEM, METHOD, AND RECORDING MEDIUM

(54) 発明の名称 : 同義語抽出システム、方法および記録媒体

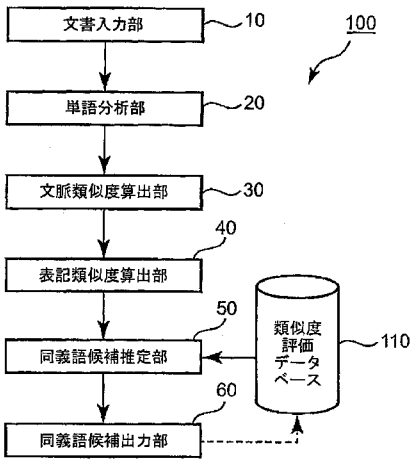

図 1

10 Text input unit
20 Term analysis unit
30 Context similarity calculation unit
40 Notation similarity calculation unit
50 Synonym candidate inference unit
60 Synonym candidate output unit
110 Similarity evaluation database

(57) **Abstract**: In order to ameliorate vagueness of a text having a synonym that is only viable in a text group pertaining to a specific matter, such as proposals, specifications, and the like pertaining to building an information system, in this synonym extraction system, by using the extraction record of an indicator of similarity to data, which can be measured with regards to a term combination and which are the term appearance count and the fraction of the count among terms, to infer and apply a similarity indicator having a high likelihood of extracting synonym candidates from a text having a synonym that is only viable in a text group pertaining to a specific matter, such as proposals, specifications, and the like pertaining to building an information system, synonyms that are only viable in a text group pertaining to a specific matter are highly precisely extracted without necessitating correct solution information or a large corpus. The synonym extraction system is provided with a text input unit, a term analysis unit, a context similarity calculation unit, a notation similarity calculation unit, a similarity evaluation database, a synonym candidate inference unit, and a synonym candidate output unit.

(57) 要約:

[続葉有]

WO 2014/002776 A1

(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, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書の曖昧さを改善するために、同義語抽出システムは、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、単語の出現個数や単語間の個数の割合といった単語組合せに関して計量可能なデータに対する類似度の指標の抽出実績を利用することによって、同義語候補を抽出する可能性の高い類似度の指標を推測して適用することで、正解情報や大量のコーパスを必要とすることなく、特定の案件に関する文書群でのみ成り立つ同義語を高精度で抽出する。同義語抽出システムは、文書入力部と、単語分析部と、文脈類似度算出部と、表記類似度算出部と、類似度評価データベースと、同義語候補推定部と、同義語候補出力部と、を備える。

明 細 書

発明の名称

同義語抽出システム、方法および記録媒体

技術分野

本発明は、同義語抽出システム、方法および記録媒体に関し、特に、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、同義語を抽出する同義語抽出システム、方法および記録媒体に関する。

背景技術

近年、情報処理装置を用いて、自然言語で書かれた文書を分析して、その文書の意味や意義を自動抽出するシステムが開発されている。そのなかで、文書中の同義語の取り扱いが問題になることがある。尚、同義語とは、意義は同じで語形が異なっている語、換言すれば、発音や表記は異なるが、意味の同じである語をいう。

同義語抽出システムに関する先行技術の一例が、特許文献１に「単語意味関係抽出装置」として記載されている。この特許文献１に開示された単語意味関係抽出装置は、テキストから抽出した単語の組に対してそれぞれ異なる複数種類の類似度を要素とする素性ベクトルを生成する手段と、既知の辞書を参照し、前記素性ベクトルに対して単語意味関係を示すラベルを付与する手段と、前記ラベルが付与された複数の素性ベクトルに基づいて単語意味関係判定ルールを学習する手段と、前記学習した単語意味関係判定ルールに基づいて、任意の単語の組に対して単語意味関係を判定する手段と、を備える。このような構成により、学習により複数の類似性の的確な統合を行い、高精度な単語意味関係抽出を行うことを実現している。

また、同義語抽出システムに関する先行技術の他の例が、特許文献２に「同義語展開システム及び同義語展開方法」として記載されている。この特許文献２に

開示された同義語展開システムは、プロセッサと、前記プロセッサに接続されるメモリと、記憶装置と、を備える少なくとも一以上の計算機を備える。前記プロセッサは、前記メモリに格納された同義語展開処理のためのプログラムを実行することによって、次に述べる処理を実行する。まず、前記プロセッサは、ある単語の係り先となる単語を含む第１文脈情報が格納された第１データベースを参照して、第１単語の前記第１文脈情報と第２単語の前記第１文脈情報とを比較することによって、前記第１単語と前記第２単語との間の意味の近さを示す類似度を計算する。そして、前記プロセッサは、前記計算された類似度が高い少なくとも一以上の前記第２単語を前記第１単語の同義語候補に決定し、前記決定された少なくとも一以上の同義語候補とその類似度とを含む第１情報を出力する。その後、前記プロセッサは、ある単語から文章中で所定の語数内に出現する単語を含む第２文脈情報が格納された第２データベースを参照して、前記第１単語の第２文脈情報と、前記出力された第１情報に含まれる少なくとも一以上の同義語候補の第２文脈情報と、を比較することによって、前記少なくとも一以上の同義語候補が出現する文脈とが一致する確率を示す文脈適合度を計算する。引き続いて、前記プロセッサは、前記少なくとも一以上の同義語候補の類似度と、前記計算された文脈適合度とに基づいて、前記同義語候補の同義語展開スコアを計算し、前記少なくとも一以上の同義語候補とその同義語展開スコアとを含む第２情報を出力する。最後に、前記プロセッサは、前記出力された第２情報に含まれる少なくとも一以上の同義語候補のうち、前記計算された同義語展開スコアの高い同義語候補を前記第１単語の同義語に決定し、前記決定された第１単語の同義語を含む第３情報を出力する。このような構成により、文書中の単語を同じ意味を表す同義語に展開する際に、その単語の出現文脈に沿った意味に展開し、文書検索、文書分類などの精度を向上させることを実現している。

さらに、同義語抽出システムに関する先行技術の他の例が、特許文献３に「辞書登録装置、辞書登録方法および辞書登録プログラム」として記載されている。この特許文献３に開示された辞書登録装置は、辞書に登録されていない単語を辞書へ登録する装置であって、単語を保持する辞書を記憶する辞書記憶手段と、入力文書を形態素解析し未知語を抽出する形態素解析部と、前記未知語の前方と後

方の少なくとも一方の単語を結合した拡張未知語を生成する未知語範囲拡張部と、前記未知語を拡張した部分の表記が一致する単語であって前記辞書に登録されている既登録単語を検索する部分一致検索部と、前記既登録単語のうち前記未知語に相当する部分の表記の文字属性と前記未知語の表記の文字属性とに基づき、表記の類似性を判定する表記類似性判定部と、前記表記類似性判例部が前記既登録単語のうち前記未知語に相当する部分の表記と前記未知語の表記とが類似すると判定した場合に、前記拡張未知語を前記辞書に登録する辞書登録部とを備える。このような構成により、複数の文字種が含まれる単語も同時に辞書に登録することができ、未知語抽出の精度を高めることができ、抽出された未知語の確認などのユーザの作業負担を軽減することができる。

先行技術文献

特許文献

特許文献１：特開２０１１－１１８５２６号公報

特許文献２：特開２０１０－２８７０２０号公報

特許文献３：特開２００６－１５５５２８号公報

発明の概要

発明が解決しようとする課題

このような先行技術の第一の課題は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、同義語の抽出に、特許文献１の先行技術による同義語の抽出方法を適用すると、特定の案件に関する文書群でのみ成り立つ同義語を抽出することができないことである。その理由は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語は、意図せずに埋め込まれており、事前にその同義関係を把握することが難しく、特許文献１の従来手法で用いられているような学習に供する正解情報としての既知の辞書を準備することが困難であるためである。

上記先行技術の第二の課題は、情報システム構築に関する提案書や仕様書等と

いった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、同義語の抽出に、上記先行技術による同義語の抽出方法を適用すると、同義語の抽出率が低くなってしまうことである。その理由は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書の多くは、文章量が限られたスモールコーパスであるため、特許文献２の従来手法で用いられているような、単語の係り先となる単語を含む第１文脈情報が格納された第１データベースや、単語から文章中で所定の語数内に出現する単語を含む第２文脈情報が格納された第２データベースといった係り受けや共起語などのコーパスが分析対象と同質なテキストデータを用意することが困難で、大量の分析対象と同質のコーパスを前提とした類似判定を行うことが難しいためである。

尚、特許文献３に開示された辞書登録装置における表記類似性判定部は、部分一致検索部が検索した単語に含まれる部分文字列のうち、未知語に相当する部分が、形態素解析部により抽出された未知語と類似するか否かを判定しているに過ぎない。

本発明の目的は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、正解情報や大量のコーパスを必要とすることなく、特定の案件に関する文書群でのみ成り立つ同義語を高精度で抽出する、同義語抽出システム、方法および記録媒体を提供することにある。

課題を解決するための手段

本発明に係る同義語抽出システムは、文書を分析して同義語を抽出する同義語抽出システムであって、対象とする文書もしくは文書群の入力を受け付ける文書入力部と；各文章に使用されている全単語の抽出および単語の定量的特徴である単語計量情報、単語の定性的特徴である単語情報の抽出を行う単語分析部と；この単語分析部で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語が使用された文脈に関する文脈情報を作成し、文脈類似度算出方法によって各単語の文脈情報間の類似性を各単語組合せの文脈類似度として算出する文脈類似度算出部と；上記単語分析部で抽出された各文章に使用されてい

る各単語について、各単語の文字構成に関する表記情報を作成し、表記類似度算出方法によって各単語の表記情報間の類似性を各単語組合せの表記類似度として算出する表記類似度算出部と；過去に同義語かどうか判定された単語組合せに関して、文書内での単語組合せの単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報、および単語間が同義語かどうかの判定結果を収集して蓄積し、単語計量情報の値によって、単語間の文脈類似度情報と単語間の表記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報を任意の類似度評価方法によって算出し、単語組合せの単語の単語計量情報について対応する類似度評価情報を応答する類似度評価データベースと；対象文書内の各単語組合せについて、上記単語分析部で抽出した各単語の単語計量情報に対応する類似度評価情報を、上記類似度評価データベースに問い合わせ、応答される類似度評価情報、および各単語間の文脈類似度と表記類似度から所定の同義判定方法によって単語類似度を算出することで、単語組合せの同義性を判定し、同義語候補の組合せとして抽出する同義語候補推定部と；同義語候補を出力する同義語候補出力部と；を備える。

発明の効果

本発明によれば、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、正解情報や大量のコーパスを必要とすることなく、特定の案件に関する文書群でのみ成り立つ同義語を高精度で抽出することが可能となる。

図面の簡単な説明

図１は本発明の一実施形態に係る同義語抽出システムの構成を示すブロック図である。

図２は図１に示した同義語抽出システムの動作例を示すシーケンス図である。

図３は本発明の第１の実施例に係る同義語抽出システムの構成を示すブロック図である。

図４は単語共起表Ｅの一部の例を示す説明図である。

図５は単語Ｓｉ間の文脈類似度Ｌｅｐｑの一部の例を示す説明図である。

図６は単語Ｓｉ間の表記類似度Ｌｗｐｑの一部の例を示す説明図である。

図７は出現数Ｐ、文脈類似度Ｌｅと表記類似度Ｌｗの蓄積データの例を示す説明図である。

図８は単語類似度Ｌｐｑを算出した結果の一部の例を示す説明図である。

発明を実施するための形態

［実施形態］

最初に、本発明の一実施形態について、図面を参照して詳細に説明する。

図１は、本発明の一実施形態に係る同義語抽出システム１００の構成を示すブロック図である。

図１を参照すると、本発明の一実施形態に係る同義語抽出システム１００は、基本的に電子機器内もしくはサーバと電子機器およびこれらを相互に接続するインターネット等の情報通信ネットワークからなるシステム内に、少なくとも、文書入力部１０、単語分析部２０、文脈類似度算出部３０、表記類似度算出部４０、同義語候補推定部５０、同義語候補出力部６０、及び類似度評価データベース１１０、を含む。

図示の同義語抽出システム１００は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、同義語を抽出する同義語抽出システムである。

少し詳細に述べると、同義語抽出システム１００は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、単語の出現個数や単語間の個数の割合といった単語組合せに関して計量可能なデータに対する類似度の指標の抽出実績を利用することによって、同義語候補を抽出する可能性の高い（同義語候補の生成パターンに応じた）類似度の指標を推測して適用することで、正解情報や大量のコーパスを必要とすることなく、特定の案件に関する文書群でのみ成り立つ同義語を高精度で抽出する、同義語抽出システムである。

電子機器で同義語抽出システムを構成する場合、同義語抽出システム１００は、プログラム制御により動作するコンピュータで実現可能である。図示はしないが、

この種のコンピュータは、周知のように、データを入力する入力装置と、データ処理装置と、データ処理装置での処理結果を出力する出力装置と、種々のデータベースとして働く補助記憶装置とを備えている。そして、データ処理装置は、プログラムを記憶するリードオンリメモリ（ＲＯＭ）と、データを一時的に記憶するワークエリアとして使用されるランダムアクセスメモリ（ＲＡＭ）と、ＲＯＭに記憶されたプログラムに従って、ＲＡＭに記憶されているデータを処理する中央処理装置（ＣＰＵ）とから構成される。

この場合、入力装置が文書入力部１０として働く。データ処理装置が、単語分析部２０、文脈類似度算出部３０、表記類似度算出部４０、および同義語候補推定部５０として働く。補助記憶装置が類似度評価データベース１１０として動作する。出力装置が同義語候補出力部６０として働く。

次に、同義語抽出システム１００を構成する各構成要素の動作について説明する。

文書入力部１０は、対象とする文書もしくは文書群の入力を受け付ける。

単語分析部２０は、文書もしくは文書群を構成する各文章に形態素解析や構文解析を適用することで、各文章に使用されている全単語の抽出および単語の定量的特徴である単語計量情報、単語の定性的特徴である単語情報の抽出を行う。

ここで、単語は名詞、動詞、形容詞など単独で意味をなす自立語に限定しても良い。また、上記単語計量情報とは、単語組合せに関して計量可能なデータであり、例えば単語組合せが使用された文書の文字数や単語数、もしくはそれぞれの単語の出現数、出現数が少ない単語側の出現数、出現数が多い単語側の出現数、単語間の出現数比率、文字数などのいずれか一つもしくはいくつかが適切である。上記単語情報は、単語の文字構成や抽出元の文を同定可能とする情報に加え、必要に応じて抽出元の文の段落や目次上の項目、単語の品詞、単語間の係り受け関係などを含めても良い。

文脈類似度算出部３０は、単語分析部２０で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語が使用された文脈に関する文脈情報を作成する。

ここで、上記文脈情報とは、各単語がどのような文脈で使用されたかを示す情

報であり、単語前後の任意範囲の文字列や、任意の共起判定方法で任意の単語と共起関係とみなされた共起語とその共起数を１文単位でまとめた共起セット、もしくは共起セットを任意の範囲の文章群について集計した共起ベクトル、などが有効である。

また、上記文脈情報の他の例としては、上記共起セットもしくは上記共起ベクトルの各共起語をシソーラスなどに基づき概念語に変換した概念セットや概念ベクトルなどの概念的な文脈情報も適している。

ここで、上記共起判定方法としては、１文、１段落内の全文章、目次上の同一項目内での全文章、文書全体など、文書の特徴に合わせて共起語と見なす範囲を設定して良く、１文内での共起する動詞、および目次上の同一項目内の文章内の名詞のように品詞毎に共起とみなす範囲を変えても良い。さらに、単語情報に単語間の係り受け関係が含まれる場合は、係り受け関係のある単語かどうかを上記共起判定方法として利用しても良い。また、共起数は共起回数でも良いが、共起回数を単語毎の全共起語数で除した頻度などでも良い。

さらに文脈類似度算出部３０は、任意に設定した文脈類似度算出方法によって各単語の文脈情報間の類似性を各単語組合せの文脈類似度として算出する。

ここで、上記文脈類似度算出方法とは、各単語の文脈情報の間の類似性を示す指標の算出方法であって、i) 上記文脈情報が単語前後の任意範囲の文字列である場合は文字列中で一致する文字の個数もしくは割合、もしくは文字列間の編集距離と単調減少の関係にある関数値を文脈類似度とする方法、ii) 上記文脈情報が共起セットの場合は共起セット内で一致した共起語の個数を文脈類似度とする方法、およびiii) 上記文脈情報が共起ベクトルの場合は共起ベクトル間のコサイン類似度や、共起ベクトル間のユークリッド距離と単調減少の関係にある関数値を文脈類似度とする方法のいずれかが適している。

表記類似度算出部４０は、単語分析部２０で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語の文字構成に関する表記情報を作成する。

ここで、上記表記情報とは、各単語がどのような表記で使用されたかを示す情報であり、単語の文字列が相当する。また、単語が複合語である場合は複合語を

構成する部分的な熟語である構成語で複合語を分解し、構成語の組合せを上記表記情報としても良い。

さらに表記類似度算出部４０は、任意に設定した表記類似度算出方法によって各単語の表記情報間の類似性を各単語組合せの表記類似度として算出する。

ここで、上記表記類似度算出方法とは、各単語の表記情報の間の類似性を示す指標の算出方法であって、i)上記表記情報が単語の文字列である場合は単語の文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を文脈類似度とする方法、および ii)上記表記情報が構成語の組合せの場合は単語間で一致した各構成語の個数もしくは割合を文脈類似度とする方法のいずれかが適している。

また、任意の加重方法で複合語内の構成語に重み付けし、より重みが大きい構成語が一致しているほど単語間の類似度が高くなるように指標を与えても良い。さらに、単語間で一致しない構成語が有る場合、その構成語間のシソーラス距離などで意味的な類似性を定量化し、一致しない構成語の意味的な類似性が高いほど、単語間の類似度が高くなるように指標を与えても良い。

類似度評価データベース１１０は、文書入力部１０で対象とした文書に限らず過去に同義語かどうか判定された単語組合せに関して、文書内での単語組合せの単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報、および単語間が同義語かどうかの判定結果を収集して蓄積し、単語計量情報の値によって、単語間の文脈類似度情報と単語間の表記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報を任意の類似度評価方法によって算出し、同義語候補推定部５０からの任意の単語組合せの単語の単語計量情報について、対応する上記類似度評価情報を応答するデータベースである。

ここで、上記文脈類似度情報は、単語の文脈情報に基づく単語間の類似性を表す情報であればよく、例えば、上記文脈類似度や上記文脈類似度に基づく単語組合せの相対順位や偏差値などが考えられる。同様に、上記表記類似度情報は、単語の表記情報に基づく単語間の類似性を表す情報であればよく、例えば、上記表記類似度や上記表記類似度に基づく単語組合せの相対順位や偏差値などが考えら

れる。

また、上記類似度評価方法は、単語の単語計量情報に関して、単語間の文脈類似度情報、単語間の表記類似度情報が同義語の判定にそれぞれどの程度、有効であるかを示す統計情報を算出可能な分析方法であればよい。例えば、上記類似度評価方法は、i) 同義語と判定された単語組合せからなる同義語セット群について、文脈類似度情報を表記類似度情報で除した値を目的変数とし、各同義語セットの単語計量情報のいくつか（例えば単語組合せで多い側の単語の出現数と、少ない側の単語の出現数）を説明変数とした重回帰分析による重回帰式を、上記類似度評価情報として算出する方法や、ii) 単語の出現数および単語間の出現数比率をそれぞれ軸とした２次元平面上に各同義語セットを配置した時に、同義性の抽出において文脈類似度情報が表記類似度情報より有効であった同義語セットの重心（例えば単語の出現数と、単語間の出現数比率の座標）である文脈類似度有効重心と、表記類似度情報が文脈類似度情報より有効であった同義語セットの重心である表記類似度有効重心を上記類似度評価情報として算出する方法などが有効である。

他にも、上記類似度評価方法は、iii) 上記単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報を前提条件とした時に、同義語と判定される条件付確率を上記類似度評価情報として算出する方法などでも良い。また、上記「単語の単語計量情報」として「単語の出現数」を想定する場合の出現数は、単語組合せ毎の単語の出現数の和でも良いし、出現数が小さい方の単語出現数もしくは出現数が大きい方の単語出現数でも良い。

同義語候補推定部５０は、対象文書内の各単語組合せについて、単語分析部２０で抽出した各単語の単語計量情報に対応する上記類似度評価情報を、類似度評価データベース１１０に問い合わせ、応答される上記類似度評価情報、および各単語間の文脈類似度と表記類似度から所定の同義判定方法によって単語類似度を算出することで、単語組合せの同義性を判定し、同義語候補の組合せとして抽出する。

ここで、上記同義判定方法は、単語計量情報から推測される、同義語の抽出により有効な類似度に基づく同義語候補の判定方法であれば良い。

例えば、上記同義判定方法は、i) 上記類似度評価情報が、上記重回帰式である場合は、上記重回帰式に各単語組合せの説明変数とした各単語計量情報（例えば、多い側の単語の出現数と、少ない側の単語の出現数）を代入し、得られる上記目的変数の値と単調増加の関係にある関数値を文脈類似度の重み付け係数に、上記目的変数の値と単調減少の関係にある関数値を表記類似度の重み付け係数にした線形和に基づく平均値を単語類似度とする方法などが有効である。

また、上記同義判定方法は、ii) 上記類似度評価情報が、上記２次元平面上における文脈類似度有効重心および表記類似度有効重心であった場合は、上記文脈類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係にある関数値を文脈類似度の係数に、上記表記類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係にある関数値を表記類似度の係数にした線形和を単語類似度とする方法なども有効である。

さらに、上記同義判定方法は、iii) 文脈類似度と表記類似度のそれぞれの上記係数を比較し、係数が大きい方の類似度のみを単語類似度とする方法なども有効である。

同義語候補出力部６０は、同義語候補推定部５０で抽出した同義語候補を出力する。

ここで、出力形態は、文書内における同義語候補の組合せを色分けや太字による強調などで明示することで、文書全体を出力する形態などが適当である。他にも、出力形態としては、同義語候補の組合せを抽出した表などの形態であって良い。また、出力形態としては、同義語候補とされた単語を主ノード、その共起語を中間ノード、概念を端ノードとして関係をリンクで結んだグラフを表示し、同義語候補とされた単語を最短で繋ぐリンクを色分けして強調するなどの形態であって良い。また、出力形態としては、同義語候補を抽出する際に用いた非類似度などで同義語間に定量的な同義度を付加し、同義度が任意に設定された閾値より大きい同義語のみに表示を限定しても良い。もしくは、出力形態としては、同義語候補間の同義度によって色分けや太字による強調もしくはグラフの単語の文字の大きさなどに強弱を与えるなどしても良い。

また、各出力形態を選択できるようにして、ベースとなる表示形態から必要に応じて表やグラフに移行できるようにしてもよい。また、必要に応じて動詞や名詞などを選択的に出力するようにしてもよい。

さらに同義語候補出力部６０は、出力した同義語候補の内、同義語と確定された単語組合せを分析者に選択させ、この単語組合せに関する単語計量情報、および各単語間の文脈類似度と表記類似度を上記類似度評価データベース１１０に登録する。

次に、図１及び図２のシーケンス図を参照して、本発明の実施形態に係る同義語抽出システム１００の全体の動作について詳細に説明する。なお、図２に示すシーケンス図及び以下の説明は処理例であり、適宜求める処理に応じて処理順等を入れ替えたり処理を戻したり繰り返したりすることを行ってもよい。

文書入力部１０は、対象とする文書もしくは文書群の入力を受け付ける（図２のステップＡ１）。

単語分析部２０は、文書もしくは文書群を構成する各文章に形態素解析や構文解析を適用することで、各文章に使用されている全単語の抽出および単語の定量的特徴である単語計量情報、単語の定性的特徴である単語情報の抽出を行う。（ステップＡ２）。

文脈類似度算出部３０は、単語分析部２０で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語が使用された文脈に関する文脈情報を作成する（ステップＡ３）。

さらに文脈類似度算出部３０は、任意に設定した文脈類似度算出方法によって各単語の文脈情報間の類似性を各単語組合せの文脈類似度として算出する（ステップＡ４）。

表記類似度算出部４０は、単語分析部２０で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語の文字構成に関する表記情報を作成する（ステップＡ５）。

さらに表記類似度算出部４０は、任意に設定した表記類似度算出方法によって各単語の表記情報間の類似性を各単語組合せの表記類似度として算出する（ステップＡ６）。

類似度評価データベース１１０は、文書入力部１０で対象とした文書に限らず過去に同義語かどうか判定された単語組合せに関して、文書内での単語組合せの単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報、および単語間が同義語かどうかの判定結果を収集して蓄積し、単語計量情報の値によって、単語間の文脈類似度情報と単語間の表記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報を任意の類似度評価方法によって算出し、同義語候補推定部５０からの任意の単語組合せの単語の単語計量情報について、対応する前記類似度評価情報を応答する（ステップＡ７）。

同義語候補推定部５０は、対象文書内の各単語組合せについて、単語分析部２０で抽出した各単語の単語計量情報に対応する上記類似度評価情報を、類似度評価データベース１１０に問い合わせ、応答される前記類似度評価情報、および各単語間の文脈類似度と表記類似度から所定の同義判定方法によって単語類似度を算出することで、単語組合せの同義性を判定し、同義語候補の組合せとして抽出（推定）する（ステップＡ８）。

同義語候補出力部６０は、同義語候補推定部５０で抽出（推定）した同義語候補を出力する（ステップＡ９）。

さらに同義語候補出力部６０は、出力した同義語候補の内、同義語と確定された単語組合せを分析者に選択させ、この単語組合せに関する各単語の単語計量情報、および各単語間の文脈類似度と表記類似度を上記類似度評価データベース１１０に登録する（ステップＡ１０）。

次に、本発明の実施形態に係る同義語抽出システム１００の効果について説明する。

本実施形態では、単語の出現個数や単語間の個数の割合といった単語組合せに関して計量可能なデータに対する類似度の抽出実績のような、文書の特徴による変化が少なく収集しやすい統計的情報を利用することによって、同義語セットであった単語組合せを抽出した確率がより高い類似度の指標を重視した同義語候補の抽出を行うように構成されている。そのため、単語の出現の頻度の偏りが大きい誤記パターン、単語の出現の頻度の偏りが小さく記載者が複数人で分担して執筆したなどで発生した用語の不統一パターン、出現頻度が少なく文脈類似度の精

度が期待できないパターン、出現頻度が多く文脈類似度が有効な情報と成るパターンといった、同義語の生成パターンに合った類似性の評価が可能になる。その結果、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から同義語を抽出できる。

尚、上記本発明の実施形態に係る同義語抽出システム１００は、同義語抽出方法として実現され得る。また、上記本発明の実施形態に係る同義語抽出システム１００は、同義語抽出プログラムによりコンピュータによって実行させるようにしても良い。

［実施例１］

次に、図３を参照して、具体的な第１の実施例を用いて、本発明の一実施形態に係る同義語抽出システム１００の動作について説明する。

本第１の実施例では、次のことを目的としている。

先ず、同義語抽出システム１００は、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書Ｄ内に含まれる特定の案件に関する文書群でのみ成り立つ同義語候補Ａを推定する。そして、同義語抽出システム１００は、推定結果を出力することで、誤字の検出や未登録の用語に関する用語集の作成や語の統一を支援する。また、本第１の実施例では、同義語抽出システム１００は、図３に示されるように、文書解析システムＹと、インターネット・サーバＺとで構成されるものとする。

文書解析システムＹは、分析実施者Ｂの持つＰＣ端末上で動作し、入力部及び出力部を介して、分析実施者Ｂが同義語を抽出したい文書群を構成する文章の入力と、同義語候補Ａの提示を実現する。

インターネット・サーバＺは、通信ネットワークを介して文書解析システムＹを実装した分析実施者Ｂの持つＰＣ端末と接続されている。インターネット・サーバＺは、文書解析システムＹからの任意の単語組合せの単語の単語計量情報に対応する上記類似度評価情報の問い合わせに対し、単語計量情報の値によって、単語間の文脈類似度情報と単語間の表記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報の検索を可能にする装置である。

図３と図１との対応関係について説明する。

文書入力部１０は、ＰＣ端末の入力部として動作する。単語分析部２０と、文脈類似度算出部３０と、表記類似度算出部４０と、同義語候補推定部５０とは、文書解析システムＹ内に含まれている。同義語候補出力部６０は、ＰＣ端末の出力部として動作する。類似度評価データベース１１０はインターネット・サーバＺ内に含まれている。

この様な手段を備えた文書解析システムＹ、インターネット・サーバＺは以下のような動作をする。

文書解析システムＹは、入力部から、分析実施者Ｂが特定の案件に関する文書から、意義は同じで語形が異なっている同義語候補Ａを推定したい文書群を構成する文書Ｄの入力を受け付ける。そして、文書解析システムＹは、文書Ｄを構成する文書の文章毎に形態素解析および構文解析を適用し、文書を構成する単語に分解し、単語毎の抽出元の文および品詞を解析することで、名詞および、動詞、形容詞、形容動詞を単語Ｗとして抽出する。なお、動詞の内でサ行変格活用に属する動詞は活用部分を除去しいわゆるサ変名詞化した形態で抽出する。

さらに文書解析システムＹは、文書Ｄに含まれる単語Ｗの内で名詞を単語Ｓとし、各単語Ｓｉ（ｉ＝１、２、・・・、ｎ）について、文書Ｄ内での出現数Ｐｉを計量する。

さらに文書解析システムＹは、文書Ｄに含まれる単語Ｗの内で名詞を単語Ｓとし、各単語Ｓｉ（ｉ＝１、２、・・・、ｎ）について、特定の単語Ｓｉと同一文中で共起関係にある名詞、動詞、形容詞を、共起語Ｖｊ（ｊ＝１、２、・・・、ｍ）として抽出し、単語Ｓｉに対する各共起語Ｖｉｊの共起回数を共起数Ｎｉｊとして集計し、全ての単語Ｓに対する各共起語Ｖについて表形式にまとめた単語共起表Ｅを作成する。なお、単語共起表Ｅの単語Ｓｉに対する各共起語Ｖｊの共起数Ｎｉｊをまとめたデータセットを単語共起ベクトルＮｉと呼ぶ。

例えば、文書Ｄの単語Ｓｉとして「交通費計算システム」、「通勤費計算」、「遅延証明」、「交通費精算サービス」、「通勤計算」などの単語が含まれていたとする。この場合、単語共起表Ｅは、図４のような、各行に単語Ｓｉを各列に共起語Ｖｉｊを配置し、その共起数Ｎｉｊを記載した表になる。また、図４の単語Ｓｉの行

のデータセットが単語共起ベクトルＮｉに相当し、「交通費計算システム」の単語共起ベクトルＮｉは｛４、２、１、１、１、０、２、０、０、・・・｝のように表される。なお、単語Ｓと共起語Ｖはいずれも名詞を含むため、先に単語として選択された単語も、他の単語が単語の場合は共起語として扱い、相互で重複して登録する。

さらに、文書解析システムＹは、同義性を評価する単語Ｓｐ（ｉ＝ｐ）と単語Ｓｑ（ｉ＝ｑ）に関して、単語Ｓｐに対応する単語共起ベクトルＮｐと単語Ｓｑに対応する単語共起ベクトルＮｑの間のコサイン類似度を文脈類似度Ｌｅｐｑとして算出する。例えば、図４の単語Ｓｉ間の文脈類似度Ｌｅｐｑの一部は、図５のような表で示される。

さらに文書解析システムＹは、単語Ｓｐおよび単語Ｓｑのそれぞれの文字列を表記情報として抽出し、文字列間の編集距離ｄｐｑを算出し、さらに単語Ｓｐおよび単語Ｓｑの文字数の内で、多い方の文字数Ｐｐｑｍａｘを算出することで、以下の数式１により単語Ｓｐおよび単語Ｓｑの表記類似度Ｌｗｐｑとして算出する。

Ｌｗｐｑ＝１－ｄｐｑ／（Ｐｐｑｍａｘ＋ｋ）　…　数式１

ここで、ｋは式中の分数の分母を０にしないための定数で０．１以下の値が適切である。例えば、編集距離の算出条件として挿入・削除・置換のコストをそれぞれ１、ｋ＝０．１として、図４の単語Ｓｉ間の表記類似度Ｌｗｐｑの一部は、図６のような表で示される。

インターネット・サーバＺは、文書Ｄに限らず過去に同義語と判定された単語組合せである同義語セットに関して、その同義語セットが使用された各文書内での各同義語の出現数Ｐを、単語間の文脈類似度Ｌｅ、および単語間の表記類似度Ｌｗを収集して蓄積する。また、インターネット・サーバＺは、収集された同義語セット群について、文脈類似度Ｌｅを単語間の表記類似度Ｌｗで除した類似度比を目的変数とし、各同義語セットの単語組合せで多い方の出現数Ｐｍａｘと、少ない方の単語の出現数Ｐｍｉｎを説明変数とした重回帰分析を行い、以下の数式２のような、単語の出現数ＰｍａｘおよびＰｍｉｎの組合せによって、単語間の文脈類似度Ｌｅと表記類似度Ｌｗがそれぞれどの程度、同義語の判定に有効に

なるかを示す統計的な関係を表す式を算出する。さらに、インターネット・サーバZは、文書解析システムYからの問い合わせに応じて、問い合わせ対象の単語組合せの出現数PmaxおよびPminに対応するLe／Lwの値を算出し、応答する。

Le／Lw＝α1×Pmax＋α2×Pmin＋β … 数式2

ここで、α1は単語の出現数Pmaxの重回帰係数、α2は単語の出現数Pminの重回帰係数、βは切片に相当する。例えば、図7のような出現数P、文脈類似度Leと表記類似度Lwの蓄積データからなる同義語セットのデータに基づく重回帰式は、以下の数式3のようになる。

Le／Lw＝0．0039×Pmax＋0．041×Pmin＋0．53 … 数式3

なお、各同義語セットの単語組合せで多い方の出現数Pmaxは文脈類似性に必要な情報量の充実性と相関することを、少ない方の単語の出現数Pminは表記類似度が近い誤字・脱字である可能性と相関することを想定しており、単語間の出現数の和や比率、文章の文字数などを説明変数に加えたり、代替するなどしても良い。

次に文書解析システムYは、上記重回帰式に単語Spおよび単語Sqの文書D中の出現数に基づく出現数PpqmaxおよびPpqminを代入し、以下の数式4のように、得られたLe／Lwの値の2乗を文脈類似度Lepqの重み付け係数に、得られたLe／Lwの値の2乗の逆数を表記脈類似度Lwpqの重み付け係数とした線形和に基づく平均値を、単語間類似度Lpqとして算出する。

Lpq＝((Le／Lw)^2×Lepq＋(Lw／Le)^2×Lwpq)／2 … 数式4

なお、上記重み付け係数は上記数式4のような連続値ではなく、得られたLe／Lwの値が1より大きい場合は文脈類似度Lepqの重み付け係数を1、表記脈類似度Lwpqの重み付け係数を0にし、得られたLe／Lwの値が1の場合は文脈類似度Lepqの重み付け係数を1／2、表記脈類似度Lwpqの重み付け係数を1／2にし、得られたLe／Lwの値が1より小さい場合は文脈類似度Lepqの重み付け係数を0、表記脈類似度Lwpqの重み付け係数を1にする

ような、不連続値を与えても良い。これは、上記数式２でＬｅ／Ｌｗが１より大きい場合は、文脈類似度Ｌｅが表記類似度Ｌｗよりも同義語の判定に有効と考えられるパターンであることを意味し、Ｌｅ／Ｌｗが１の場合は、同義語の判定の有効性が文脈類似度Ｌｅと表記類似度Ｌｗとで同等であるパターンであることを意味し、Ｌｅ／Ｌｗが１より小さい場合は、表記類似度Ｌｗが文脈類似度Ｌｅよりも同義語の判定に有効と考えられるパターンであることを意味するためである。

さらに、文書解析システムＹは、単語類似度Ｌｐｑが任意の判定閾値Ｔより大きい単語Ｓｐと単語Ｓｑの組合せを、単語の共起ベクトルの意味的な類似性が高く、同義語の可能性が想定される単語の組合せである同義語候補Ａとして抽出する。この処理を全ての単語Ｓｉの組合せについて行う。

例えば、図４～図７の例で、「交通費計算システム」と「交通費精算サービス」、「通勤費計算」と「通勤計算」、「遅延証明」と「通勤費計算」の組合せの単語類似度Ｌｐｑを算出した結果は、図８の表のようになる。判定閾値Ｔ＝０．７５とすると、文脈類似度Ｌｅと表記類似度Ｌｗの単純平均ではいずれも判定閾値Ｔを越える組合せは無いが、上記数４に基づいて単語類似度Ｌを算出した結果は、「交通費計算システム」と「交通費精算サービス」、「通勤費計算」と「通勤計算」が判定閾値Ｔより大きく、この文章内では同義語である可能性があると判定される。これは、単語の出現数がある程度多く文脈類似性が有効かつ、両単語とも極端に少ない出現数では無く誤字・脱字とは考えにくい「交通費計算システム」と「交通費精算サービス」の単語類似度Ｌは文脈類似度に近い値となり、逆に単語の出現数がある程度多く文脈類似性が有効だが、一方の単語の出現数が極端に少なく誤字・脱字の可能性が高い「通勤費計算」と「通勤計算」の単語類似度Ｌは表記類似度に近い値となるような重み付けが統計情報により付与されたためである。このように単語の出現数などの計量可能なデータに対する類似度の指標の抽出実績を利用することで、有効な類似度を重視した同義語候補の抽出ができ、的確な同義語の検出が可能になる。

さらに文書解析システムＹは、同義語候補Ａａ｛Ｓｐ、Ｓｑ｝について、要求文書Ｄで該当する同義語候補Ａａ｛Ｓｐ、Ｓｑ｝を色分けもしくは太字による強調などの加工を行い、加工後の要求文書Ｄを、出力部から出力する。

以上説明したように、本発明の同義語抽出システムによれば、情報システム構築に関する提案書や仕様書等といった、特定の案件に関する文書群でのみ成り立つ同義語のある文書から、正解情報や大量のコーパスを必要とすることなく、特定の案件に関する文書群でのみ成り立つ同義語を高精度で抽出することが可能となり、誤解に基づく混乱や失敗などの削減につなげられることにある。その理由は、単語の出現個数や単語間の個数の割合といった単語組合せに関して計量可能なデータに対する類似度の指標の抽出実績のような、文書の特徴による変化が少なく収集しやすい統計的情報を利用することによって、同義語候補を抽出する可能性の高い（同義語候補の生成パターンに応じた）類似度の指標を推測して適用することで、同義語の生成パターンに応じた類似度の指標を適用した単語間の類似性算出を可能にしているためである。

以上、実施形態（実施例）を参照して本願発明を説明したが、本願発明は上記実施形態（及び実施例）に限定されるものではない。本願発明の構成や詳細には、本願発明のスコープ内で当業者が理解し得る様々な変更をすることができる。

産業上の利用可能性

本発明によれば、ソフトウェアやシステムの開発における要件定義などの作業においてやり取りされる各種文書に関して、文書の曖昧さに繋がる同義語を除外することで文書の理解・作成・修正を支援することが可能になり、手戻りの減少や顧客満足の向上などシステム開発の効率化に関する用途に適用できる。また、同義語を精度良く抽出できるので、翻訳システムに用いて訳し分けに利用できる。

符号の説明

１０　文書入力部
２０　単語分析部
３０　文脈類似度算出部
４０　表記類似度算出部
５０　同義語候補推定部
６０　同義語候補出力部

１００　同義語抽出システム

１１０　類似度評価データベース

Ｄ　文書

Ａ　同義語

Ｙ　文書解析システム

Ｚ　インターネット・サーバ

この出願は、２０１２年６月２５日に出願された、日本特許出願第２０１２－１４１６８０号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。

請 求 の 範 囲

［請求項１］

文書を分析して同義語を抽出する同義語抽出システムであって、

対象とする文書もしくは文書群の入力を受け付ける文書入力部と、

各文章に使用されている全単語の抽出および単語の定量的特徴である単語計量情報、単語の定性的特徴である単語情報の抽出を行う単語分析部と、

前記単語分析部で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語が使用された文脈に関する文脈情報を作成し、文脈類似度算出方法によって各単語の文脈情報間の類似性を各単語組合せの文脈類似度として算出する文脈類似度算出部と、

前記単語分析部で抽出された各文章に使用されている各単語について、各単語の文字構成に関する表記情報を作成し、表記類似度算出方法によって各単語の表記情報間の類似性を各単語組合せの表記類似度として算出する表記類似度算出部と、

過去に同義語かどうか判定された単語組合せに関して、文書内での単語組合せの単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報、および単語間が同義語かどうかの判定結果を収集して蓄積し、前記単語計量情報の値によって、前記単語間の文脈類似度情報と前記単語間の表記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報を類似度評価方法によって算出し、単語組合せの単語の単語計量情報について対応する前記類似度評価情報を応答する類似度評価データベースと、

対象文書内の各単語組合せについて、前記単語分析部で抽出した各単語の単語計量情報に対応する前記類似度評価情報を、前記類似度評価データベースに問い合わせ、応答される前記類似度評価情報、および各単語間の文脈類似度と表記類似度から所定の同義判定方法によって単語類似度を算出することで、単語組合せの同義性を判定し、同義語候補の組合せとして抽出する同義語候補推定部と、

前記同義語候補を出力する同義語候補出力部と、

を備えたことを特徴とする同義語抽出システム。

［請求項２］

前記単語計量情報は、単語組合せに関して計量可能なデータであって、

i) 単語組合せが使用された文書の文字数や単語数、

ii) それぞれの単語の出現数、

iii) 出現数が少ない単語側の出現数、

iv) 出現数が多い単語側の出現数、

v) 単語間の出現数比率、および

vi) 文字数

のいずれか一つもしくはいくつかである、ことを特徴とする請求項１に記載の同義語抽出システム。

［請求項３］

前記文脈情報は、各単語がどのような文脈で使用されたかを示す情報であって、

i) 単語前後の任意範囲の文字列、

ii) 共起判定方法で単語と共起関係とみなされた共起語とその共起数を１文単位でまとめた共起セット、

iii) 該共起セットを所定の範囲の文章群について集計した共起ベクトル、および

iv) 前記共起セットもしくは前記共起ベクトルの各共起語をシソーラスに基づき概念語に変換した概念セットや概念ベクトル

のグループから選択されたいずれか１つである、ことを特徴とする請求項１又は２に記載の同義語抽出システム。

［請求項４］

前記文脈類似度算出方法は、各単語の文脈情報の間の類似性を示す指標の算出方法であって、

i) 前記文脈情報が単語前後の任意範囲の文字列である場合は文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を文脈類似度とする方法、

ii) 前記文脈情報が共起セットの場合は共起セット内で一致した共起語の個数を文脈類似度とする方法、および

iii) 前記文脈情報が共起ベクトルの場合は共起ベクトル間のコサイン類似度や、共起ベクトル間のユークリッド距離と単調減少の関係にある関数値を文脈類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項３に記載の同義語抽出システム。

［請求項５］

前記表記情報は、各単語がどのような表記で使用されたかを示す情報であって、

i) 単語の文字列、および

ii) 単語が複合語である場合は複合語を構成する構成語の組合せ

のグループから選択されたいずれか１つである、ことを特徴とする請求項１乃至４のいずれか１項に記載の同義語抽出システム。

［請求項６］

前記表記類似度算出方法は、各単語の表記情報の間の類似性を示す指標の算出方法であって、

i) 前記表記情報が単語の文字列である場合は文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を表記類似度とする方法、および

ii) 前記表記情報が構成語の組合せの場合は単語間で一致した各構成語の個数もしくは割合を表記類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項５に記載の同義語抽出システム。

［請求項７］

前記文脈類似度情報は、単語の文脈情報に基づく単語間の類似性を表す情報であって、前記文脈類似度や前記文脈類似度に基づく単語組合せの相対順位や偏差値であり、

前記表記類似度情報は、単語の表記情報に基づく単語間の類似性を表す情報であって、前記表記類似度や前記表記類似度に基づく単語組合せの相対順位や偏差値である、

ことを特徴とする請求項１乃至６のいずれか１項に記載の同義語抽出システム。

［請求項８］

前記類似度評価方法は、単語の単語計量情報に関して、単語間の文脈類似度情報、単語間の表記類似度情報が同義語の判定にそれぞれどの程度、有効であるかを示す統計情報を算出可能な分析方法であって、

i) 同義語と判定された単語組合せからなる同義語セット群について、文脈類似度情報を表記類似度情報で除した値を目的変数とし、各同義語セットの単語計量情報のいくつかを説明変数とした重回帰分析による重回帰式を、前記類似度評価情報として算出す方法する方法、

ii) 単語の出現数および単語間の出現数比率をそれぞれ軸とした２次元平面上に各同義語セットを配置した時に、同義性の抽出において文脈類似度情報が表記類似度情報より有効であった同義語セットの重心である文脈類似度有効重心と、表記類似度情報が文脈類似度情報より有効であった同義語セットの重心である表記類似度有効重心を前記類似度評価情報として算出する方法、および

iii) 前記単語計量情報、前記単語間の文脈類似度情報、前記単語間の表記類似度情報を前提条件とした時に、同義語と判定される条件付確率を前記類似度評価情報として算出する方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項７に記載の同義語抽出システム。

［請求項９］

前記同義判定方法は、単語計量情報から推測される、同義語の抽出により有効な類似度に基づく同義語候補の判定方法であって、

i) 前記類似度評価情報が、前記重回帰式である場合は、前記重回帰式に各単語組合せの説明変数とした各単語計量情報を代入し、得られる前記目的変数の値と単調増加の関係にある関数値を文脈類似度の重み付け係数に、前記目的変数の値と単調減少の関係にある関数値を表記類似度の重み付け係数にした線形和に基づく平均値を単語類似度とする方法、

ii) 前記類似度評価情報が、前記２次元平面上における文脈類似度有効重心および表記類似度有効重心であった場合は、前記文脈類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係

にある関数値を文脈類似度の係数に、前記表記類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係にある関数値を表記類似度の係数にした線形和を単語類似度とする方法、および

iii) 文脈類似度と表記類似度のそれぞれの前記係数を比較し、係数が大きい方の類似度のみを単語類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項８に記載の同義語抽出システム。

［請求項１０］

前記同義語候補出力部は、出力した同義語候補の内、同義語と確定された単語組合せを分析者に選択させ、この単語組合せに関する単語計量情報、および各単語間の文脈類似度と表記類似度を前記類似度評価データベースに登録する、

ことを特徴とする請求項１乃至９のいずれか１項に記載の同義語抽出システム。

［請求項１１］

文書を分析して同義語を抽出する同義語抽出方法であって、

対象とする文書もしくは文書群の入力を受け付ける文書受付工程と、

各文章に使用されている全単語の抽出および単語の定量的特徴である単語計量情報、単語の定性的特徴である単語情報の抽出を行う単語情報抽出工程と、

前記単語情報抽出工程で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語が使用された文脈に関する文脈情報を作成し、文脈類似度算出方法によって各単語の文脈情報間の類似性を各単語組合せの文脈類似度として算出する文脈類似度算出工程と、

前記単語情報抽出工程で抽出された各文章に使用されている各単語について、各単語の文字構成に関する表記情報を作成し、表記類似度算出方法によって各単語の表記情報間の類似性を各単語組合せの表記類似度として算出する表記類似度算出工程と、

過去に同義語かどうか判定された単語組合せに関して、文書内での単語組合せの単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報、および単語間が同義語かどうかの判定結果を収集して蓄積する類似度評価データベースに、前記単語計量情報の値によって、前記単語間の文脈類似度情報と前記単語間の表

記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報を類似度評価方法によって算出させ、単語組合せの単語の単語計量情報について対応する前記類似度評価情報を応答させる工程と、

対象文書内の各単語組合せについて、前記単語情報抽出工程で抽出した各単語の単語計量情報に対応する前記類似度評価情報を、前記類似度評価データベースに問い合わせ、応答される前記類似度評価情報、および各単語間の文脈類似度と表記類似度から所定の同義判定方法によって単語類似度を算出することで、単語組合せの同義性を判定し、同義語候補の組合せとして抽出する同義語候補推定工程と、

前記同義語候補を出力する同義語候補出力工程と、

を含むことを特徴とする同義語抽出方法。

[請求項１２]

前記単語計量情報は、単語組合せに関して計量可能なデータであって、

i) 単語組合せが使用された文書の文字数や単語数、

ii) それぞれの単語の出現数、

iii) 出現数が少ない単語側の出現数、

iv) 出現数が多い単語側の出現数、

v) 単語間の出現数比率、および

vi) 文字数

のいずれか一つもしくはいくつかである、ことを特徴とする請求項１１に記載の同義語抽出方法。

[請求項１３]

前記文脈情報は、各単語がどのような文脈で使用されたかを示す情報であって、

i) 単語前後の任意範囲の文字列、

ii) 共起判定方法で単語と共起関係とみなされた共起語とその共起数を１文単位でまとめた共起セット、

iii) 該共起セットを所定の範囲の文章群について集計した共起ベクトル、および

iv) 前記共起セットもしくは前記共起ベクトルの各共起語をシソーラスに基づ

き概念語に変換した概念セットや概念ベクトル
のグループから選択されたいずれか１つである、ことを特徴とする請求項１１又は１２に記載の同義語抽出方法。

［請求項１４］

前記文脈類似度算出方法は、各単語の文脈情報の間の類似性を示す指標の算出方法であって、

ｉ）前記文脈情報が単語前後の任意範囲の文字列である場合は文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を文脈類似度とする方法、

ii）前記文脈情報が共起セットの場合は共起セット内で一致した共起語の個数を文脈類似度とする方法、および

iii）前記文脈情報が共起ベクトルの場合は共起ベクトル間のコサイン類似度や、共起ベクトル間のユークリッド距離と単調減少の関係にある関数値を文脈類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項１３に記載の同義語抽出方法。

［請求項１５］

前記表記情報は、各単語がどのような表記で使用されたかを示す情報であって、

ｉ）単語の文字列、および

ii）単語が複合語である場合は複合語を構成する構成語の組合せ

のグループから選択されたいずれか１つである、ことを特徴とする請求項１１乃至１４のいずれか１項に記載の同義語抽出方法。

［請求項１６］

前記表記類似度算出方法は、各単語の表記情報の間の類似性を示す指標の算出方法であって、

i）前記表記情報が単語の文字列である場合は文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を表記類似度とする方法、および

ii）前記表記情報が構成語の組合せの場合は単語間で一致した各構成語の個数

もしくは割合を表記類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項１５に記載の同義語抽出方法。

［請求項１７］

前記文脈類似度情報は、単語の文脈情報に基づく単語間の類似性を表す情報であって、前記文脈類似度や前記文脈類似度に基づく単語組合せの相対順位や偏差値であり、

前記表記類似度情報は、単語の表記情報に基づく単語間の類似性を表す情報であって、前記表記類似度や前記表記類似度に基づく単語組合せの相対順位や偏差値である、

ことを特徴とする請求項１１乃至１６のいずれか１項に記載の同義語抽出方法。

［請求項１８］

前記類似度評価方法は、単語の単語計量情報に関して、単語間の文脈類似度情報、単語間の表記類似度情報が同義語の判定にそれぞれどの程度、有効であるかを示す統計情報を算出可能な分析方法であって、

i) 同義語と判定された単語組合せからなる同義語セット群について、文脈類似度情報を表記類似度情報で除した値を目的変数とし、各同義語セットの単語計量情報のいくつかを説明変数とした重回帰分析による重回帰式を、前記類似度評価情報として算出す方法する方法、

ii) 単語の出現数および単語間の出現数比率をそれぞれ軸とした２次元平面上に各同義語セットを配置した時に、同義性の抽出において文脈類似度情報が表記類似度情報より有効であった同義語セットの重心である文脈類似度有効重心と、表記類似度情報が文脈類似度情報より有効であった同義語セットの重心である表記類似度有効重心を前記類似度評価情報として算出する方法、および

iii) 前記単語計量情報、前記単語間の文脈類似度情報、前記単語間の表記類似度情報を前提条件とした時に、同義語と判定される条件付確率を前記類似度評価情報として算出する方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項１７に記載の同義語抽出方法。

[請求項１９]

前記同義判定方法は、単語計量情報から推測される、同義語の抽出により有効な類似度に基づく同義語候補の判定方法であって、

i) 前記類似度評価情報が、前記重回帰式である場合は、前記重回帰式に各単語組合せの説明変数とした各単語計量情報を代入し、得られる前記目的変数の値と単調増加の関係にある関数値を文脈類似度の重み付け係数に、前記目的変数の値と単調減少の関係にある関数値を表記類似度の重み付け係数にした線形和に基づく平均値を単語類似度とする方法、

ii) 前記類似度評価情報が、前記２次元平面上における文脈類似度有効重心および表記類似度有効重心であった場合は、前記文脈類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係にある関数値を文脈類似度の係数に、前記表記類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係にある関数値を表記類似度の係数にした線形和を単語類似度とする方法、および

iii) 文脈類似度と表記類似度のそれぞれの前記係数を比較し、係数が大きい方の類似度のみを単語類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項１８に記載の同義語抽出方法。

[請求項２０]

前記同義語候補出力工程は、出力した同義語候補の内、同義語と確定された単語組合せを分析者に選択させ、この単語組合せに関する単語計量情報、および各単語間の文脈類似度と表記類似度を前記類似度評価データベースに登録する、

ことを特徴とする請求項１１乃至１９のいずれか１項に記載の同義語抽出方法。

[請求項２１]

コンピュータに文書を分析させて、同義語を抽出させる同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体であって、前記コンピュータに、

対象とする文書もしくは文書群の入力を受け付ける文書受付手順と、

各文章に使用されている全単語の抽出および単語の定量的特徴である単語計量情報、単語の定性的特徴である単語情報の抽出を行う単語情報抽出手順と、

前記単語情報抽出手順で抽出された各文章に使用されている各単語について、単語情報を利用して、各単語が使用された文脈に関する文脈情報を作成し、文脈類似度算出方法によって各単語の文脈情報間の類似性を各単語組合せの文脈類似度として算出する文脈類似度算出手順と、

前記単語情報抽出手順で抽出された各文章に使用されている各単語について、各単語の文字構成に関する表記情報を作成し、表記類似度算出方法によって各単語の表記情報間の類似性を各単語組合せの表記類似度として算出する表記類似度算出手順と、

過去に同義語かどうか判定された単語組合せに関して、文書内での単語組合せの単語計量情報、単語間の文脈類似度情報、単語間の表記類似度情報、および単語間が同義語かどうかの判定結果を収集して蓄積する類似度評価データベースに、前記単語計量情報の値によって、前記単語間の文脈類似度情報と前記単語間の表記類似度情報がそれぞれどの程度、同義語の判定に有効になるかを示す統計情報である類似度評価情報を類似度評価方法によって算出させ、単語組合せの単語の単語計量情報について対応する前記類似度評価情報を応答させる手順と、

対象文書内の各単語組合せについて、前記単語情報抽出手順で抽出した各単語の単語計量情報に対応する前記類似度評価情報を、前記類似度評価データベースに問い合わせ、応答される前記類似度評価情報、および各単語間の文脈類似度と表記類似度から所定の同義判定方法によって単語類似度を算出することで、単語組合せの同義性を判定し、同義語候補の組合せとして抽出する同義語候補推定手順と、

前記同義語候補を出力する同義語候補出力手順と、

を実行させる同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２２］

前記単語計量情報は、単語組合せに関して計量可能なデータであって、

i) 単語組合せが使用された文書の文字数や単語数、

ii) それぞれの単語の出現数、

iii) 出現数が少ない単語側の出現数、

iv) 出現数が多い単語側の出現数、

v) 単語間の出現数比率、および

vi) 文字数

のいずれか一つもしくはいくつかである、ことを特徴とする請求項２１に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２３］

前記文脈情報は、各単語がどのような文脈で使用されたかを示す情報であって、

i) 単語前後の任意範囲の文字列、

ii) 共起判定方法で単語と共起関係とみなされた共起語とその共起数を１文単位でまとめた共起セット、

iii) 該共起セットを所定の範囲の文章群について集計した共起ベクトル、および

iv) 前記共起セットもしくは前記共起ベクトルの各共起語をシソーラスに基づき概念語に変換した概念セットや概念ベクトル

のグループから選択されたいずれか１つである、ことを特徴とする請求項２１又は２２に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２４］

前記文脈類似度算出方法は、各単語の文脈情報の間の類似性を示す指標の算出方法であって、

i) 前記文脈情報が単語前後の任意範囲の文字列である場合は文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を文脈類似度とする方法、

ii) 前記文脈情報が共起セットの場合は共起セット内で一致した共起語の個数を文脈類似度とする方法、および

iii) 前記文脈情報が共起ベクトルの場合は共起ベクトル間のコサイン類似度や、共起ベクトル間のユークリッド距離と単調減少の関係にある関数値を文脈類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項２３に

記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２５］

前記表記情報は、各単語がどのような表記で使用されたかを示す情報であって、

i) 単語の文字列、および

ii) 単語が複合語である場合は複合語を構成する構成語の組合せ

のグループから選択されたいずれか１つである、ことを特徴とする請求項２１乃至２４のいずれか１項に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２６］

前記表記類似度算出方法は、各単語の表記情報の間の類似性を示す指標の算出方法であって、

i) 前記表記情報が単語の文字列である場合は文字列中で一致する文字の個数もしくは割合や、文字列間の編集距離と単調減少の関係にある関数値を表記類似度とする方法、および

ii) 前記表記情報が構成語の組合せの場合は単語間で一致した各構成語の個数もしくは割合を表記類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項２５に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２７］

前記文脈類似度情報は、単語の文脈情報に基づく単語間の類似性を表す情報であって、前記文脈類似度や前記文脈類似度に基づく単語組合せの相対順位や偏差値であり、

前記表記類似度情報は、単語の表記情報に基づく単語間の類似性を表す情報であって、前記表記類似度や前記表記類似度に基づく単語組合せの相対順位や偏差値である、

ことを特徴とする請求項２１乃至２６のいずれか１項に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２８］

前記類似度評価方法は、単語の単語計量情報に関して、単語間の文脈類似度情

報、単語間の表記類似度情報が同義語の判定にそれぞれどの程度、有効であるかを示す統計情報を算出可能な分析方法であって、

i) 同義語と判定された単語組合せからなる同義語セット群について、文脈類似度情報を表記類似度情報で除した値を目的変数とし、各同義語セットの単語計量情報のいくつかを説明変数とした重回帰分析による重回帰式を、前記類似度評価情報として算出す方法する方法、

ii) 単語の出現数および単語間の出現数比率をそれぞれ軸とした２次元平面上に各同義語セットを配置した時に、同義性の抽出において文脈類似度情報が表記類似度情報より有効であった同義語セットの重心である文脈類似度有効重心と、表記類似度情報が文脈類似度情報より有効であった同義語セットの重心である表記類似度有効重心を前記類似度評価情報として算出する方法、および

iii) 前記単語計量情報、前記単語間の文脈類似度情報、前記単語間の表記類似度情報を前提条件とした時に、同義語と判定される条件付確率を前記類似度評価情報として算出する方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項２７に記載の同義語プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項２９］

前記同義判定方法は、単語計量情報から推測される、同義語の抽出により有効な類似度に基づく同義語候補の判定方法であって、

i) 前記類似度評価情報が、前記重回帰式である場合は、前記重回帰式に各単語組合せの説明変数とした各単語計量情報を代入し、得られる前記目的変数の値と単調増加の関係にある関数値を文脈類似度の重み付け係数に、前記目的変数の値と単調減少の関係にある関数値を表記類似度の重み付け係数にした線形和に基づく平均値を単語類似度とする方法、

ii) 前記類似度評価情報が、前記２次元平面上における文脈類似度有効重心および表記類似度有効重心であった場合は、前記文脈類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係にある関数値を文脈類似度の係数に、前記表記類似度有効重心と、各単語の出現数と各単語間の出現数比率からなる座標のユークリッド距離と単調減少の関係に

ある関数値を表記類似度の係数にした線形和を単語類似度とする方法、および

iii) 文脈類似度と表記類似度のそれぞれの前記係数を比較し、係数が大きい方の類似度のみを単語類似度とする方法

のグループから選択されたいずれか１つである、ことを特徴とする請求項２８に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

［請求項３０］

前記同義語候補出力手順は、出力した同義語候補の内、同義語と確定された単語組合せを分析者に選択させ、この単語組合せに関する単語計量情報、および各単語間の文脈類似度と表記類似度を前記類似度評価データベースに登録する、

ことを特徴とする請求項２１乃至２９のいずれか１項に記載の同義語抽出プログラムを記録したコンピュータ読み取り可能な記録媒体。

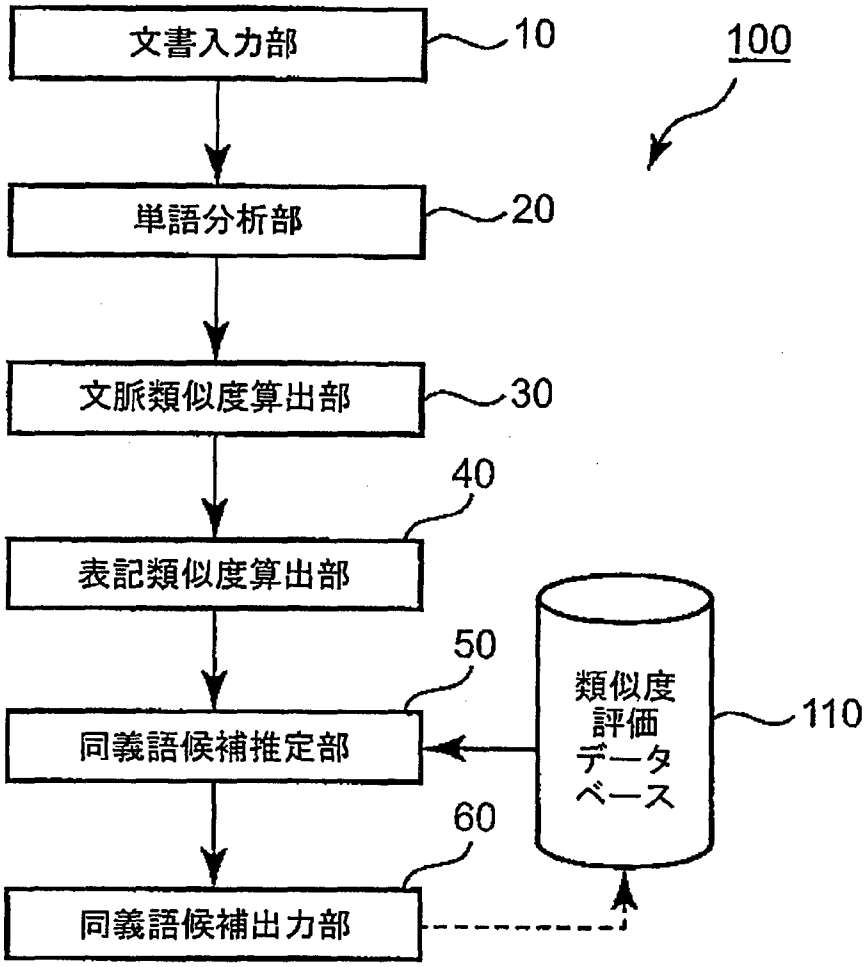
文書入力部
10
100
単語分析部
20
文脈類似度算出部
30
40
表記類似度算出部
50
同義語候補推定部
類似度
評価
データ
ベース
110
60
同義語候補出力部

図 1

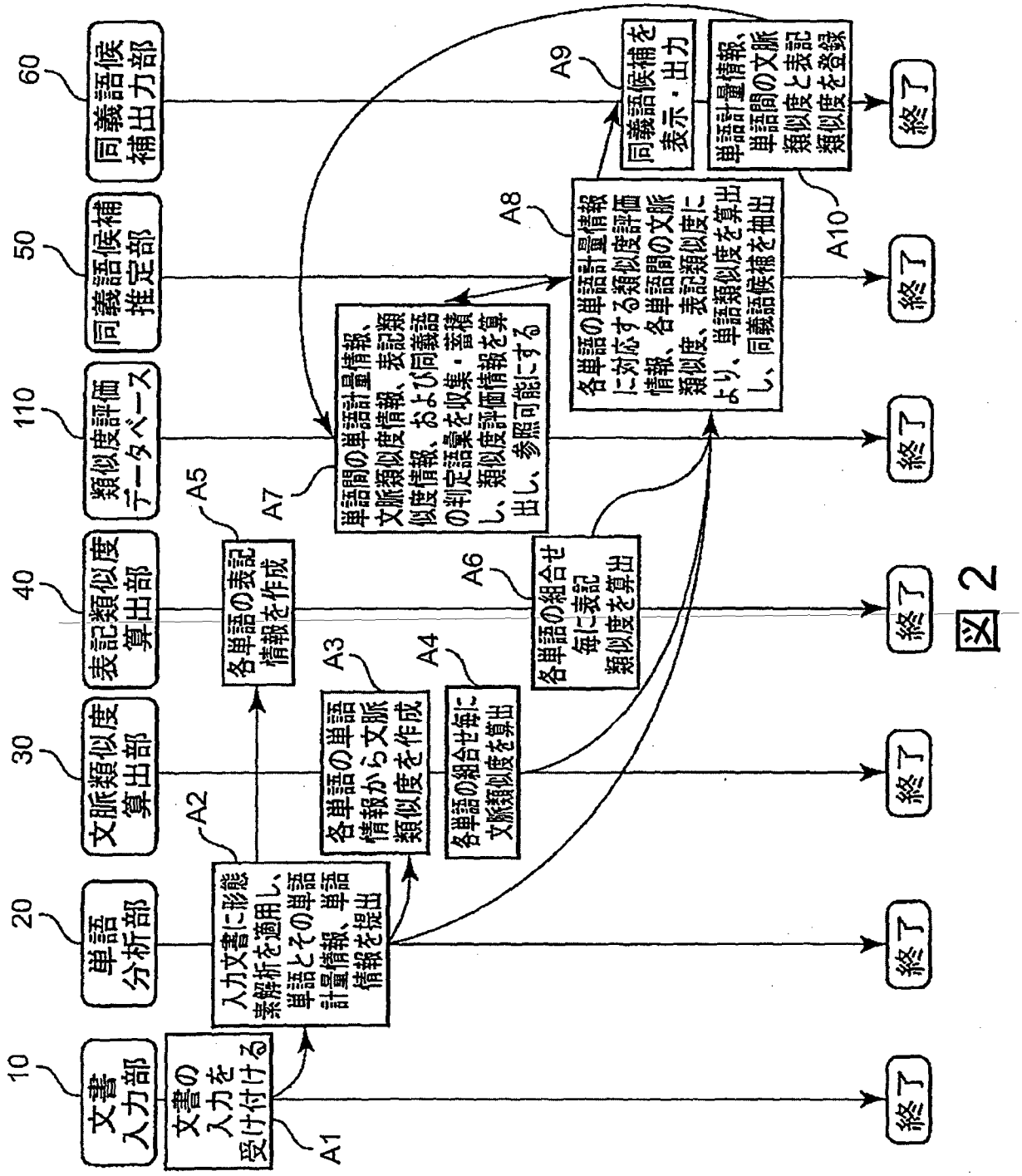
10
文書入力部
20
単語分析部
30
文脈類似度算出部
40
表記類似度算出部
110
類似度評価データベース
50
同義語候補推定部
60
同義語候補出力部
文書の入力を受け付ける
A1
入力文書に形態素解析を適用し、単語とその単語計量情報、単語情報を提出
A2
各単語の単語情報から文脈類似度を作成
A3
各単語の組合せ毎に文脈類似度を算出
A4
各単語の表記情報を作成
A5
各単語の組合せ毎に表記類似度を算出
A6
単語間の単語計量情報、文脈類似度情報、表記類似度情報、および同義語の判定語彙を収集・蓄積し、類似度評価情報を算出し、参照可能にする
A7
各単語の単語計量情報に対応する類似度評価情報、各単語間の文脈類似度、表記類似度により、単語類似度を算出し、同義語候補を抽出
A8
同義語候補を表示・出力
A9
単語計量情報、単語間の文脈類似度と表記類似度を登録
A10
終了

図2

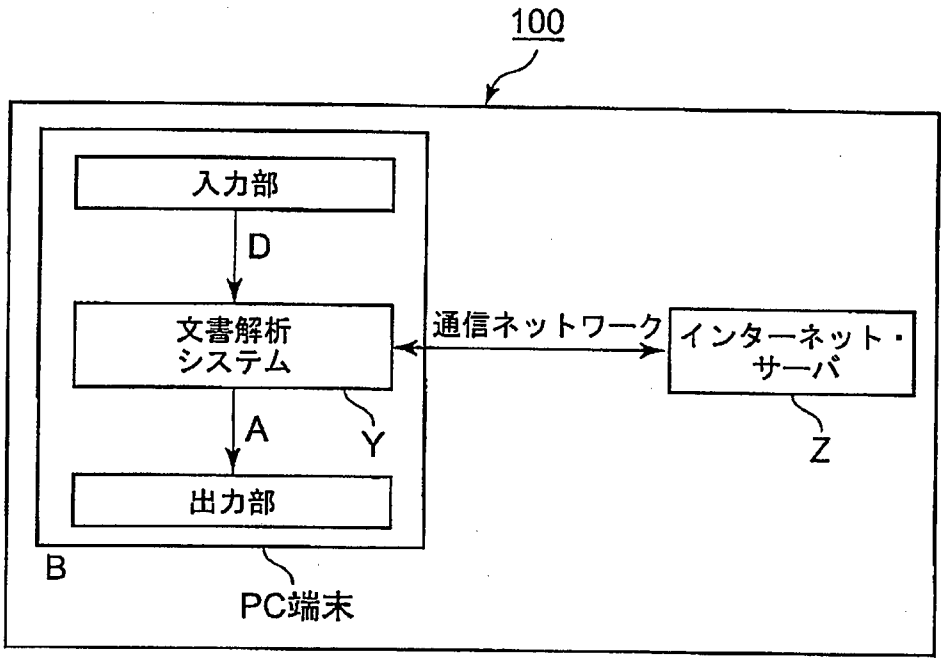

図 3

単語Si	出現数Fi	共起語Vi									
		利用	ウィンドウ	並べて	表示	方法	機能	以下	両面	構築	…
交通費計算システム	35	4	2	1	1	1	0	2	0	0	…
通勤費計算	51	0	0	0	1	0	2	1	1	0	…
遅延証明	21	0	0	0	1	0	1	1	1	1	…
交通費精算システム	43	5	2	1	1	1	0	0	0	1	…
通勤計算	2	0	0	0	1	0	3	1	2	1	…
－	－	－	－	－	－	－	－	－	－	－	…

図 4

単語組合せ		文脈類似度 Lepq
単語Sp	単語Sq	
交通費計算システム	交通費精算サービス	0.87
通勤費計算	通勤計算	0.32
遅延証明	通勤費計算	0.73
⋮	⋮	⋮

図 5

単語組合せ				編集距離	表記類似度
単語Sp	出現数 Pp	単語Sq	出現数 Pq	dpq	Lwpq
交通費計算システム	9	交通費精算サービス	9	4	0.56
通勤費計算	5	通勤計算	4	1	0.80
遅延証明	4	通勤費計算	5	5	0.02
⋮	⋮	⋮	⋮	⋮	⋮

図 6

同義語セット				大きい方の単語出現数	小さい方の単語出現数	文脈類似度	表記類似度	類似度比
同義語A	単語出現数	同義語A'	単語出現数	Pmax	Pmin	Le	Lw	Le／Lw
競争的資金	81	競争資金	3	81	3	0.643	0.800	0.80
出荷管理票	78	出荷記録	17	78	17	0.821	0.400	2.05
勤務実績	76	出勤データ	33	76	33	0.818	0.200	4.09
勤怠管理	68	就業管理	14	68	14	0.758	0.500	1.52
同義語組	63	同義セット	42	63	42	0.841	0.400	2.10
検索エリア	61	検案欄	26	61	26	0.655	0.400	1.64
就業時間	58	勤務時間	45	58	45	0.887	0.500	1.77
間接費	56	一般管理費	23	56	23	0.616	0.200	3.08
稼動ログ	55	稼動実績	79	79	55	0.796	0.500	1.59
在庫調整量	64	入荷制限量	28	64	28	0.761	0.200	3.81
宛先情報	52	住所情報	47	52	47	0.820	0.500	1.64
勘定ツール	49	会計システム	42	49	42	0.784	0.167	4.71
無停電装置	45	停電対策1	35	45	35	0.636	0.400	1.59
共通DB	42	汎用DB	38	42	38	0.741	0.500	1.48
共起ベクトル	42	共時ベクトル	2	42	2	0.193	0.833	0.23
日次売上集計	32	一日売上計	76	76	32	0.746	0.667	1.12
転居情報	31	移動情報	62	62	31	0.594	0.500	1.19
バックアップ電源	31	無停電装置	27	31	27	0.518	0.125	4.15
共起データ	29	共起ベクトル	48	48	29	0.518	0.333	1.56
飲食業	11	外食産業	77	77	11	0.625	0.500	1.25
同義判定ルール	7	同義評価ルール	13	13	7	0.295	0.714	0.41
間接経費	4	間接費	66	66	4	0.296	0.750	0.39
日次売上集	2	日次売上集計	32	32	2	0.126	0.833	0.15
無停電源	1	無停電電源	25	25	1	0.009	0.800	0.01

図 7

単語組合せ					大きい方の単語出現数	小さい方の単語出現数	Le／Lw	文脈類似度	表記類似度	類似度の相加平均	単語類似度
単語Sp	出現数 Pp	単語Sq	出現数 Pq		Ppqmax	Ppqmin		Lepq	Lwpq		Lpq
交通費計算システム	35	交通費精算サービス	43		43	35	1.24	0.87	0.56	0.72	0.85
通勤費計算	51	通勤計算	2		51	2	0.76	0.32	0.80	0.56	0.79
遅延証明	21	通勤費計算	51		51	21	1.05	0.73	0.02	0.37	0.41
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・		・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・

図 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2013/066286

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30(2006.01)i, *G06F17/27*(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, G06F17/27

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 1996-2013
Kokai Jitsuyo Shinan Koho 1971-2013 Toroku Jitsuyo Shinan Koho 1994-2013

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2010-152561 A (Toshiba Corp.), 08 July 2010 (08.07.2010), entire text; all drawings (Family: none)	1-30
A	JP 2009-129323 A (Hitachi, Ltd.), 11 June 2009 (11.06.2009), entire text; all drawings (Family: none)	1-30
A	Eiji HIRAO, "Development of the detection method of synonyms in the requirements documents", The Institute of Electronics, Information and Communication Engineers 2012 Nen Sogo Taikai Koen Ronbunshu, Joho System 1, 06 March 2012 (06.03.2012), page 26	1-30

☒ 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
31 July, 2013 (31.07.13)	13 August, 2013 (13.08.13)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2013/066286

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Minoru YOSHIDA, "Application of text mining", The Journal of Information Science and Technology Association, 01 June 2010 (01.06. 2010), vol.60, no.6, pages 230 to 235	1-30

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１３／０６６２８６

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, G06F17/27(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, G06F17/27

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	１９７１－２０１３年
日本国実用新案登録公報	１９９６－２０１３年
日本国登録実用新案公報	１９９４－２０１３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）

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	関連する請求項の番号
A	JP 2010-152561 A (株式会社東芝) 2010.07.08, 全文, 全図 (ファミリーなし)	１－３０
A	JP 2009-129323 A (株式会社日立製作所) 2009.06.11, 全文, 全図 (ファミリーなし)	１－３０
A	平尾　英司, 要求文書中の同義語推定手法の開発, 電子情報通信学会２０１２年総合大会講演論文集　情報・システム１, 2012.03.06, ２６ページ	１－３０

☑　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	国際調査報告の発送日
３１．０７．２０１３	１３．０８．２０１３

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　３６５９
吉田　誠
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	関連する請求項の番号
Ａ	吉田　稔，テキストマイニングの活用，情報の科学と技術，2010.06.01，Ｖｏｌ．６０　Ｎｏ．６，２３０－２３５ページ	１－３０