

明 細 書

発明の名称：表記ゆれ検出装置及び表記ゆれ検出プログラム

技術分野

[0001] 本発明の実施形態は、表記ゆれ検出装置及び表記ゆれ検出プログラムに関する。

背景技術

[0002] 一般に、同じ概念に対して複数の表記（単語）が存在していることを表記ゆれという。この表記ゆれが文書中に存在すると、ユーザが文書を検索する際や、文書から特定の用語を抽出する際などに、同じ概念の用語であるにも関わらず、適切に抽出されない場合がある。

[0003] ここで、表記ゆれに関し、これまでに種々の技術がある。例えば、対象となる文書から表記ゆれ候補と考えられる文字列を予め選別して辞書を作成し、この辞書に基づいて表記ゆれの候補の文字列を検出する方法もある。

[0004] しかしながら、この方法では予め表記ゆれの候補を手で選別するなどして辞書を作成する必要があるため、効率が悪いという欠点があった。

先行技術文献

特許文献

[0005] 特許文献1：特開2005-35288号公報

特許文献2：特開2006-53866号公報

発明の概要

発明が解決しようとする課題

[0006] 本発明が解決しようとする課題は、精度よく表記ゆれ候補を検出する表記ゆれ検出装置を提供することである。

課題を解決するための手段

[0007] 実施形態の表記ゆれ検出装置は、文書データから用語を抽出する用語抽出部と、抽出された用語の任意のペアの類似度を算出する類似度算出部と、類似度に基づいて、用語のペアが表記ゆれ候補であることを判定する表記ゆれ候

補判定部と、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いて表記ゆれ候補をグループ分けするグループ分類部と、を備える。

図面の簡単な説明

[0008] [図1]実施形態の表記ゆれ検出装置の全体構成を示すブロック図。

[図2]実施形態の表記ゆれ検出装置の表記ゆれ検出処理の一例を示すフローチャート。

[図3]実施形態の表記ゆれ検出装置の文書データ記憶部に格納される文書データの一例を示す模式図。

[図4]実施形態の表記ゆれ検出装置の用語抽出部に抽出された用語の一例を示す模式図。

[図5]実施形態の表記ゆれ検出装置の類似度算出部による類似度算出処理の一例を示すフローチャート。

[図6]実施形態の表記ゆれ検出装置の類似度算出部によって算出された類似度の一例を示す模式図。

[図7]実施形態の表記ゆれ検出装置の類似度算出部によって算出された類似度の一例を示す模式図。

[図8]実施形態の表記ゆれ検出装置の類似度重み付け部による重み付け情報の一例を示す模式図。

[図9]実施形態の表記ゆれ検出装置の表記ゆれ候補の一例を示す模式図。

[図10]実施形態の表記ゆれ検出装置のグループ解析部のグループ解析処理の一例を示す模式図。

[図11]実施形態の表記ゆれ検出装置の表記ゆれ候補の一例を示す模式図。

[図12]実施形態の表記ゆれ検出装置の出力部の一例を示す模式図。

発明を実施するための形態

[0009] 以下、本発明の実施形態について図面を参照して説明する。

[0010] 図1は本実施形態の表記ゆれ検出装置100の全体構成を示すブロック図である。図1に示すように、表記ゆれ検出装置100は、文書入力部1、用語抽出部2、類似度算出部3、重み付け部4、表記ゆれ候補判定部5、グル

ープ解析部 6、表記ゆれ種類判定部 7、出力データ生成部 8、文書データ記憶部 10、および重み付け情報記憶部 11 を備える。

[0011] 表記ゆれを検出する対象の文書データは、文書入力部 1 によって入力される。文書入力部 1 は例えばキーボードやマウスであり、文書データ記憶部 10 に格納された文書データから検出対象の文書データを選択する。文書データ記憶部 10 に記憶されている文書データは、当該文書データの種類と対応付けて記憶されている。文書データの種類とは、例えば、「契約書」、「規程」、「法令文書」、「ニュース記事」などがある。

[0012] すなわち、文書入力部 1 から、表記ゆれ検出対象の文書データと当該文書データの種類のとが入力される。なお、文書入力部 1 から直接に文書データと当該文書データの種類のとが入力されてもよい。

[0013] 用語抽出部 2 は、入力された文書データから用語（ここでは単語や複合語を意味する）の抽出を行う。本実施形態の用語抽出部 2 は、複数の抽出方法で用語を抽出する。

[0014] 類似度算出部 3 は、抽出された用語群の中の 2 つの用語の類似度を算出する。なお、類似度の算出は、編集距離に基づいて行う。編集距離とは、2 つの文字列がどの程度異なっているかを示す数値である。また、編集距離及び類似度の算出については後述する。

[0015] 重み付け部 4 は、重み付け情報記憶部 11 に格納された重み付け情報に基づいて、類似度算出部 3 で算出された類似度に文書データの種類ごとの重み付けをする重み付け処理を行う。なお、重み付け処理の詳細については後述する。

[0016] 表記ゆれ候補判定部 5 は、重み付け後の類似度に基づいて、抽出された用語群の中の 2 つの用語が表記ゆれ候補であるかを判定する。グループ解析部 6 は、判定された表記ゆれ候補群の共通の文字列と類似度とに基づき、表記ゆれ候補のグループを解析して分類する。

[0017] 表記ゆれ種類判定部 7 は、各表記ゆれの候補の表記ゆれの種類を判定し、表記ゆれ判定処理を行う。表記ゆれ検出部が検出した表記ゆれ候補は、出力

部 8 から出力される。この出力部 8 は、例えば液晶ディスプレイなどの表示装置である。

[0018] ここで、図 2 乃至図 12 を参照して表記ゆれ検出装置 100 における表記ゆれ検出処理について説明する。図 2 は表記ゆれ検出装置 100 の表記ゆれ検出処理の一例を示すフローチャートである。

[0019] まず、ユーザが、表記ゆれ検出装置 100 の文書データ入力部 1 を用いて、文書データ記憶部 10 から表記ゆれ検出対象の文書データを入力する（ステップ S10）。図 3 に、入力される文書データの一部である文書データ 101 を示す。図 3 に示すように文書データ 101 は文書種類 102 と文書データ本体 103 とを含む。文書データ 101 の文書種類 102 は、ここでの説明で一例として「規程」を挙げる。文書データ 101 が入力されると、用語抽出部 2 は、文書データ 101 から用語の抽出を行う（ステップ S20）。用語抽出部 2 の用語抽出は、例えば形態素解析や字面解析などの方法で行う。ここでは、用語抽出部 2 は、2 種類の抽出方法「抽出方法 A」および「抽出方法 B」によって、それぞれ用語抽出を行ったとする。

[0020] 図 4 に、用語抽出部 2 による用語抽出結果を示すように、用語抽出部 2 は、所定の抽出法で用語とその用語の抽出箇所を示す文書データの行番号とを対応付けて抽出する。

[0021] 図 4 の（a）は、抽出手法 A で抽出された用語を示す抽出用語テーブル 201 であり、図 4 の（b）は抽出手法 B で抽出された用語を示す抽出用語テーブル 202 である。抽出用語テーブルは 201 および 202 は、用語 ID カラム 203、用語カラム 204、行番号カラム 205、抽出手法カラム 206 を含む。用語 ID 203 には、抽出用語の識別子が格納される。用語カラム 204 には、用語抽出部 2 によって抽出された用語が格納される。行番号 205 には用語カラム 204 が抽出された行番号が格納される。抽出手法カラム 206 には、用語抽出部 2 によって用語を抽出した際に抽出書法が格納される。なお、用語抽出の際に、抽出する用語毎に抽出元の文書名や品詞の情報などが付与されてもよい。

[0022] 続いて、類似度算出部3は、抽出された用語から任意の2つの用語を選択し、選択した2つの用語の類似度を算出する（ステップS30）。ここで、図5を参照して類似度算出部3の類似度算出処理の一例について説明する。

[0023] まず、類似度算出部3では用語抽出部2の抽出結果から2つの用語が選択される（ステップS31）。なお、用語抽出部2が複数の抽出方法で用語を抽出した場合は、同じ用語抽出方法で抽出された用語の中から2つを選択する。

[0024] 類似度算出部3は、選択した2つの用語の編集距離を算出する（ステップS32）。編集距離は2つの文字列がどの程度異なるかを示す数値であり、例えば2つ用語の一方を編集して他方の用語と一致させるためにかかる操作の回数に基づいて算出される。

[0025] 操作とは、例えば、削除・置換・挿入などが挙げられる。これらの操作1回あたりのコストをあらかじめ「1」として設定しておき、一方の用語に操作を行い、他方の用語と同一の用語になるまでのコストを合計したものを編集距離とする。

[0026] なお、本実施形態では、操作の種類ごとのコストを設定して編集距離を算出する。例えば、置換の操作のうち、表記ゆれの可能性が高い、「全角／半角」、「平仮名／片仮名」、「大文字／小文字」などの文字種置換のコストを「0.1」とする。このように、操作ごとのコストをあらかじめ設定することにより、表記ゆれ検出の精度を向上することが可能である。

[0027] 類似度算出部3は、算出した編集距離を用いて2つの用語の類似度を算出する（ステップS33）。類似度は、例えば2つの用語の文字列の長さの和に対する編集距離の割合を算出し、算出した割合を1から引くことで算出される。

[0028] 例えば図4の（a）に示すテーブルの用語IDカラム203が3の用語（以下、用語3という）と用語IDカラム203が5の用語（以下、用語5という）の類似度を算出する。用語3と用語5を同一の用語にするために用語3を編集する場合、用語3の「PL」の「P（全角）」を「P（半角）」に置

換し、「L（全角）」を「L（半角）」に置換する。すなわち操作回数は「全角／半角」置換が2回であり、置換のコストを「0.1」とすると、編集距離は「0.2」である。

[0029] 用語3と用語5の文字列の長さの合計は「4」であるため、用語3と用語5の類似度は、「0.95」となる。

[0030] 類似度を算出していない用語が、用語抽出部2に抽出された用語の中にある場合（ステップS34がNo）、ステップS31に戻り、類似度算出処理を繰り返す。すべての用語の類似度を算出した場合（ステップS34がYes）、類似度算出処理を終了する。

[0031] 図6及び図7に、類似度算出部3が算出した類似度の一例を示す。ここでは、図4に示した抽出用語テーブル201と抽出用語テーブル202のテーブルに基づいて類似度を算出したとする。

[0032] 図6及び図7は、類似度を算出した用語ペアごとの類似度を示す類似度テーブルを示す図である。図6は図4の（a）に示した抽出用語テーブル201に基づいて類似度を算出した類似度テーブル301である。図7は、図4の（b）に示した抽出用語テーブル202に基づいて類似度を算出した類似度テーブル302である。類似度テーブル301および302には、用語ペアIDカラム303、抽出手法カラム304、用語Aカラム305、用語Bカラム306、用語A行番号カラム307、用語B行番号カラム308、類似度カラム309が含まれる。用語ペアIDカラム303は、類似度算出対象の用語ペアの識別子が格納される。抽出手法カラム304は抽出手法カラム206の内容が格納される。用語Aカラム305には、類似度算出対象の用語ペアの片方の用語が格納される。用語Bカラム306には、類似度算出対象の用語ペアの他方の用語が格納される。用語A行番号カラム307には、用語Aカラム305に格納された用語の抽出元の行番号カラム205の内容が格納される。用語B行番号カラム308には、用語Bカラム306に格納された用語の抽出元の行番号カラム205の内容が格納される。類似度カラム309には、類似度算出部3が算出した類似度が格納される。

- [0033] ここで、図2の説明に戻る。類似度が算出されると、類似度重み付け部4は、表記ゆれ検出対象の文書の種類と、類似度を算出した用語の抽出方法とに基づいて、類似度に重み付けを行う（ステップS40）。
- [0034] 図8を参照して、類似度重み付け部4による類似度重み付け処理について説明する。
- [0035] 図8は類似度重み付け情報テーブル111である。類似度重み付け情報テーブル111は、文書の種類112および類似度を算出した用語の抽出方法113ごとの重み付け情報が格納されている。この重み付け情報はあらかじめシステム設計者やユーザによって設定されている。本実施例では、文書の種類112として「法令文書」、「ニュース記事」、「新聞」、「規程」、及び「学術文書」が挙げられている。また、用語抽出の手法113として、「抽出手法A」、及び「抽出手法B」が挙げられている。
- [0036] すなわち、図2のステップS40で、類似度重み付け部4は、表記ゆれ検出対象の文書データの文書種類102と、類似度を算出した用語の抽出方法とに基づいて、類似度重み付け情報テーブル111を参照し、類似度に重み付けを行う。
- [0037] 続いて、表記ゆれ候補判定部5は、重み付けされた用語ペアごとの類似度に基づいて、表記ゆれ候補を判定する（ステップS50）。表記ゆれ候補の判定は、例えば類似度があらかじめ設定された閾値以上の用語ペアを抽出するなどの方法で行う。本実施形態は、類似度の閾値を0.6とする。
- [0038] ここで、図9の表記ゆれ候補テーブル401に、表記ゆれ候補を示す。図9の表記ゆれ候補テーブル401は、類似度テーブル301および302から表記ゆれ候補判定部7によって判定された表記ゆれ候補を示すものであり、類似度テーブル301および302に含まれる項目に表記ゆれ候補の識別子である表記ゆれ候補IDカラム402を追加したテーブルである。重み付け部4によって類似度に重み付けが行われたため、図6及び図7に示した類似度テーブルに含まれる用語ペアのうち、類似度が0.6を超える用語ペアは「抽出手法A」の「用語ペアID11」の用語ペアと、「抽出手法B」の

「用語ペア I D 2, 4, 6, 9」の用語ペアの5つである。

- [0039] 表記ゆれ候補が判定されると、グループ解析部6は、表記ゆれ候補の共通文字列や類似度を参照して、表記ゆれ候補のグループ分けを行う（ステップS60）。
- [0040] グループ解析部6における表記ゆれ候補のグループ分け処理は、選択した任意の2つの表記ゆれ候補XとYに含まれる用語をそれぞれ比較することによって行われる。グループ解析部6は、表記ゆれ候補Xに含まれる用語のうち少なくとも一方の用語の文字列が、表記ゆれ候補Yに含まれる用語の文字列に含まれる場合、表記ゆれ候補XおよびYを同一のグループと判別し、同一のグループIDを付与する。
- [0041] 図10は、グループ解析部6によるグループ解析処理の一例を示すフローチャートである。なお、表記ゆれ候補に付与するグループIDを1以上の整数であるnとし、グループ解析処理開始時はn=1であるとする。
- [0042] 図10に示すように、グループ解析部6は、まず、表記ゆれ候補判定部5によって判定された表記ゆれ候補の中から任意の表記ゆれ候補を選択する（ステップS61）。ここで選択した表記ゆれ候補を基準候補Xとする。また、基準候補X以外の表記ゆれ候補を、被解析候補という。
- [0043] 選択した表記ゆれ候補XにグループIDが付与されている場合（ステップS62がNo）、グループ解析部6は、ステップS61に戻り、再び表記ゆれ候補を選択する。
- [0044] 選択した表記ゆれ候補XにグループIDが付与されていない場合（ステップS62がYes）、グループ解析部6は、被解析候補の中から表記ゆれ候補を選択する（ステップS63）。ここで選択された表記ゆれ候補を被解析候補Yとする。
- [0045] 基準候補Xに含まれる用語Aと用語Bのどちらか一方が、表記ゆれ候補Yに含まれる場合（ステップS64がYes）、グループ解析部6は、非判別候補YにグループIDが付与されているかどうかを判定する（ステップS65）。なお、基準候補Xに含まれる用語Aと用語Bのどちらも、判別候補Y

に含まれない場合（ステップS 6 4がN o）、グループ解析部6はステップS 6 3に戻り、判別候補から候補を再度選択する。

[0046] 非判別候補YにグループIDが付与されていない場合（ステップS 6 5がY e s）、基準候補Xと被解析候補Yとは同一のグループであると判定し、候補Xと候補YのグループIDを「n」とする（ステップS 6 6）。

[0047] グループIDを付与すると、グループ解析部6は $n = n + 1$ とし（ステップS 6 7）、ステップS 6 8に進む。

[0048] 非判別候補YにグループIDが付与されている場合（ステップS 6 5がN o）、基準候補Xに被解析候補Yとは同一のグループIDを付与する（ステップS 7 0）。その後、ステップS 6 8に進む。

[0049] グループ解析処理を行われていない被解析候補が存在する場合（ステップS 6 8がN o）、ステップS 6 3に戻り、再び被解析候補を選択する。なお、グループ解析処理を行われていない被解析候補が存在するかどうかの判定は、例えば「候補総数-1」とグループID付与済みの候補総数を比較することにより行う。具体的には、「候補総数-1」がグループID付与済みの候補総数未満である場合、グループ解析部6は、すべての被解析候補にグループ解析処理を行っているとは判定する。逆に、「候補総数-1」がグループID付与済みの候補総数以上である場合、グループ解析部6は、グループ解析処理を行われていない被解析候補が存在すると判定する。

[0050] グループ解析処理を行われていない被解析候補が存在しない場合（ステップS 6 8がY e s）、グループ解析部6は、すべての表記ゆれ候補にグループ判定処理を行ったか否かを判定する（ステップS 6 9）。グループ解析処理を行われていない被解析候補が存在しない場合（ステップS 6 9がN o）、ステップS 6 1にもどり、表記ゆれ候補から基準候補Xを選択し処理を繰り返す。すべての表記ゆれ候補にグループ判定処理を行った場合（ステップS 6 9がY e s）、すなわち、すべての表記ゆれ候補にグループIDが付与された場合グループ解析処理を終了する。

[0051] 上述のように、本実施形態のグループ解析部6は、表記ゆれ候補をグルー

分けする。また、グループ解析部6は、すでにグループIDが付与された表記ゆれ候補に基づいてグループ間係を抽出するため、関連する表記ゆれ候補を効率よく抽出することが可能である。

[0052] 続いて、図2の説明に戻る。表記ゆれ種類判定部7は、表記ゆれ候補の種類を判定する(ステップS80)。表記ゆれ種類判定部7によって判定される表記ゆれ種類は、例えば、「スぺース違い」、「半角全角違い」、「後方一致」、「前方一致」、「片仮名平仮名違い」、「一文字違い」、「複数文字違い」である。

[0053] 「スぺース違い」は、表記ゆれ候補に含まれる用語それぞれのスぺースを削除した場合に同一の用語になる。「半角全角違い」は、表記ゆれ候補に含まれる用語それぞれを全角もしくは半角に統一した場合に同一の用語になる。「後方一致」は、表記ゆれ候補に含まれる用語の後方の文字が一致する状態をいう。「前方一致」は、表記ゆれ候補に含まれる用語のそれぞれの前方の文字が一致する状態をいう。「片仮名平仮名違い」は、表記ゆれ候補に含まれる用語それぞれを平仮名もしくは片仮名に統一した場合に同一の用語になる。「一文字違い」は、表記ゆれ候補に含まれる用語が一文字違いである。「複数文字違い」は、表記ゆれ候補に含まれる用語が複数文字違いである。

[0054] 図11に、グループ解析部6によるグループ解析処理の結果、および表記ゆれ種類分類部7による表記ゆれ種類分類処理の結果の一例を示す。図11は、表記ゆれ候補テーブル401に、グループ解析部6によるグループ解析処理の結果が格納される。表記ゆれ種類分類部7による表記ゆれ種類分類処理の結果が格納される。図12は、本実施形態の表記ゆれ検出処理後の、出力部8の一例を示す図である。図12に示すように、出色部8には表記ゆれ検出画面601が表示される。表記ゆれ検出画面601には、ソートボタン602と、ソート設定

ボタン603と、表記ゆれ検出結果表示領域604を有する。表記ゆれ検出結果表示領域604には、検出結果として、用語Aカラム305、用語Bカラム306、用語A行番号カラム307、用語B行番号カラム308、類似度カラム309、グループカラム502、表記ゆれ種類カラム503が表示される。ソートボタン602は、ソート設定ボタン603の設定に基づいて表記ゆれ検出結果表示領域604に表示された表記ゆれ検出結果の表示順序の並べ替えを行う。ここでは、ソート設定ボタン603が「グループ順」と設定されているため、グループカラム502に格納されたグループIDが小さい順に表示されている。

[0057] 上述のように、本実施形態の表記ゆれ検出装置は、あらかじめ辞書を作成することなく、表記ゆれ候補を検出することが可能である。また、編集距離の算出の際に文字種置換操作を加え、文字種置換操作をその他の操作による編集距離よりも短くすることにより、精度の高い表記ゆれ候補検出を可能とする。また、文書データの種類毎に重み付けを行うことにより、表記ゆれ候補検出の精度を向上する。

[0058] また、本実施形態の表記ゆれ検出装置は、検出した表記ゆれ候補のグループ関係を解析し、グループID毎に出力部8から出力することによって、表記ゆれの候補をユーザが効率的に確認することを可能にする。同様に、表記ゆれ候補の種類の分類毎に表示部に表示することも可能である。

[0059] なお、本実施形態の表記ゆれ検出装置100は、用語を登録する辞書記憶部を備えてもよい。この場合、辞書記憶部には所定の文書データから抽出された表記ゆれ候補を登録する。類似度算出部5は、用語抽出部2が抽出した用語と、辞書記憶部に登録された用語との類似度を算出する。これにより、社内規程など、同じ語句が使用されられると思われる文書データから効率よく表記ゆれを検出することが可能となる。

[0060] また、辞書記憶部に除外条件を登録し、除外条件を満たす表記ゆれ候補を表記ゆれ候補から削除するようにしてもよい。上記の除外条件とは、例えば、「先頭に『各』という文字があり、『各』に続く文字列が一致する用語ペ

ア」、「末尾に『等』という文字があり『等』の前の文字列が一致する用語ペア」、「末尾に『書』という文字があり、『書』の前の文字列が一致する用語ペア」などの条件である。

[0061] また、用語抽出部2が文書データから用語抽出を行う際に、用語を抽出した行番号と用語が何文字目の文字かを示す位置情報を抽出した用語に付与しても良い。これにより、同一の位置から抽出された用語が表記ゆれ候補にならないため、表記ゆれ候補の検出精度を向上することが可能である。

[0062] 以上、本発明の実施形態を説明したが、この実施形態はあくまでも例として提示したものであり、発明の範囲を限定することは意図していない。この新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。この実施形態やその変形は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。

符号の説明

[0063] 1…文書データ入力部、2…用語抽出部、3…類似度算出部、4…類似度重み付け部、5…表記ゆれ候補判定部、6…グループ関係抽出部、7…表記ゆれ種類判定部、10…文書データ記憶部、11…重み付け情報記憶部

請求の範囲

- [請求項1] 文書データから用語を抽出する用語抽出部と、
前記抽出された用語の任意のペアの類似度を算出する類似度算出部と、
前記類似度に基づいて、前記用語のペアが表記ゆれ候補であるかを判定する表記ゆれ候補判定部と、
前記表記ゆれ候補である前記用語のペアに共通に含まれる文字列に基づいて前記表記ゆれ候補をグループ分けするグループ分類部と、
を備える表記ゆれ検出装置。
- [請求項2] 文書データから用語を抽出する用語抽出部と、
前記抽出された用語の任意のペアに対して文字種置換を含む操作を行い、前記操作の回数に基づいて編集距離を算出し、前記編集距離に基づいて類似度を算出する類似度算出部と、
前記類似度に基づき、前記用語のペアが表記ゆれ候補であるかを判定する表記ゆれ候補判定部と、
を備える表記ゆれ検出装置。
- [請求項3] 前記類似度算出部は、前記抽出された用語のペアの一方の用語に置換、削除、挿入もしくは文字種置換の操作を繰り返し適用することで編集距離を求め、類似度を算出する請求項1に記載の表記ゆれ検出装置
- [請求項4] 前記表記ゆれ候補である前記用語のペアに共通に含まれる文字列に基づいて前記表記ゆれ候補をグループ分けするグループ分類部をさらに備える請求項2に記載の表記ゆれ検出装置。
- [請求項5] 文書データから用語を抽出する用語抽出部と、前記抽出された用語の任意のペアを用いて類似度を算出する類似度算出部と、前記類似度に基づいて、前記用語のペアが表記ゆれ候補であるかを判定する表記ゆれ候補判定部と、前記表記ゆれ候補である前記用語のペアに共通に含まれる文字列に基づいて前記表記ゆれ候補をグループ分けするグル

ープ分類部と、を備える表記ゆれ検出装置に用いられる表記ゆれ検出プログラムであって、

前記表記ゆれ検出装置に対して、

前記用語抽出部が文書データから用語を抽出する機能と、

前記類似度算出部が前記抽出された用語の任意のペアを用いて類似度を算出する機能と、

前記表記ゆれ候補判定部が前記類似度に基づいて、前記用語のペアが表記ゆれ候補であるかを判定する機能と、

前記グループ分類部が前記表記ゆれ候補である用語のペアに含まれる共通文字列に基づいて前記表記ゆれ候補をグループ分けする機能と、

を実現させる表記ゆれ検出プログラム。

[請求項6]

文書データから用語を抽出する用語抽出部と、前記抽出された用語の任意のペアに対して文字種置換を含む操作を行い、前記操作の回数に基づいて編集距離を算出し、前記編集距離に基づいて類似度を算出する類似度算出部と、前記類似度に基づき、前記用語のペアが表記ゆれ候補であるかを判定する表記ゆれ候補判定部と、を備える表記ゆれ検出装置に用いられる表記ゆれ検出プログラムであって、

前記表記ゆれ検出装置に対して

前記用語抽出部が文書データから用語を抽出する機能と、

前記類似度算出部が前記抽出された用語の任意のペアに対して文字種置換を含む操作を行い、前記操作の回数に基づいて編集距離を算出し、前記編集距離に基づいて類似度を算出する機能と、

前記表記ゆれ候補判定部が前記の類似度に基づき、前記用語の任ペアが表記ゆれ候補であるかを判定する機能と、

を実現させる表記ゆれ検出プログラム。