(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6427466号
(P6427466)
(45)発行日　平成30年11月21日(2018.11.21)　(24)登録日　平成30年11月2日(2018.11.2)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 320D
G06F 17/30 170A
G06F 17/30 350C

請求項の数 5　(全 11 頁)

(21)出願番号　特願2015-106871(P2015-106871)
(22)出願日　平成27年5月26日(2015.5.26)
(65)公開番号　特開2016-224482(P2016-224482A)
(43)公開日　平成28年12月28日(2016.12.28)
審査請求日　平成29年6月21日(2017.6.21)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　斉藤　いつみ
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　貞光　九月
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　浅野　久子
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】同義語ペア獲得装置、方法、及びプログラム

(57)【特許請求の範囲】
【請求項１】
文書から、正規表記語である単語分割候補、及び前記正規表記語に対して揺らいだ表記の候補である崩れ表記語である単語分割候補を含む複数の単語分割候補を生成する単語分割候補生成部と、
前記単語分割候補生成部により生成された前記複数の単語分割候補に基づいて、前記複数の単語分割候補の各々について、単語の意味ベクトルを計算する意味ベクトル計算部と、
正規表記語である前記単語分割候補の各々について、前記意味ベクトルに基づいて計算される意味類似度と、単語の読みに基づいて計算される音類似度とに基づいて、前記複数の単語分割候補をフィルタリングし、フィルタリングされた前記複数の単語分割候補から、予め定められた前記正規表記語と同一の表記であって、前記同一の表記の前記正規表記語と同一の品詞である前記単語分割候補を除いて選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得する同義語ペア獲得部と、
を含む同義語ペア獲得装置。
【請求項２】
前記同義語ペア獲得部は、
選択された前記単語分割候補の各々について、前記意味類似度と前記音類似度とに基づいて、前記複数の単語分割候補をフィルタリングし、フィルタリングされた前記複数の単

語分割候補から、予め定められた前記正規表記語と同一の表記であって、前記同一の表記の前記正規表記語と同一の品詞である前記単語分割候補を除いて更に選択し、正規表記語である前記単語分割候補と、更に選択された前記単語分割候補とのペアを、同義語ペアとして獲得する請求項１に記載の同義語ペア獲得装置。

【請求項３】

単語分割候補生成部が、文書から、正規表記語である単語分割候補、及び前記正規表記語に対して揺らいだ表記の候補である崩れ表記語である単語分割候補を含む複数の単語分割候補を生成するステップと、

意味ベクトル計算部が、前記単語分割候補生成部により生成された前記複数の単語分割候補に基づいて、前記複数の単語分割候補の各々について、単語の意味ベクトルを計算するステップと、

同義語ペア獲得部が、正規表記語である前記単語分割候補の各々について、前記意味ベクトルに基づいて計算される意味類似度と、単語の読みに基づいて計算される音類似度とに基づいて、前記複数の単語分割候補をフィルタリングし、フィルタリングされた前記複数の単語分割候補から、予め定められた前記正規表記語と同一の表記であって、前記同一の表記の前記正規表記語と同一の品詞である前記単語分割候補を除いて選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得するステップと、

を含む同義語ペア獲得方法。

【請求項４】

前記同義語ペア獲得部が獲得するステップは、

選択された前記単語分割候補の各々について、前記意味類似度と前記音類似度とに基づいて、前記複数の単語分割候補をフィルタリングし、フィルタリングされた前記複数の単語分割候補から、予め定められた前記正規表記語と同一の表記であって、前記同一の表記の前記正規表記語と同一の品詞である前記単語分割候補を除いて更に選択し、正規表記語である前記単語分割候補と、更に選択された前記単語分割候補とのペアを、同義語ペアとして獲得する請求項３に記載の同義語ペア獲得方法。

【請求項５】

コンピュータを、請求項１又は請求項２に記載の同義語ペア獲得装置の各部として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、同義語ペア獲得装置、方法、及びプログラムに係り、特に、同義語ペアを獲得するための同義語ペア獲得装置、方法、及びプログラムに関する。

【背景技術】

【０００２】

従来より、正規表記語に対して揺らいだ表記である崩れ表記語を獲得するための手法が提案されている。教師データを用いた手法としては、非特許文献１及び非特許文献２に記載されている手法が挙げられる。

【０００３】

教師データを用いない手法としては、非特許文献３及び非特許文献４に記載されている手法が挙げられる。

【先行技術文献】

【非特許文献】

【０００４】

【非特許文献１】岡崎直観，辻井潤一，“アライメント識別モデルを用いた略語定義の自動獲得”．言語処理学会第14回年次大会（NLP2008），pp．139-142

【非特許文献２】藤沼祥成，横野光，相澤彰子，“Twitter（Ｒ）上の「おはよう」を例とした崩れた表記の検出と分析．” 第27 回人工知能学会全国大会，2013.06

【非特許文献３】増山毅司，関根聡，“大規模コーパスからのカタカナ語の表記の揺れリストの自動構築”，言語処理学会第14回年次大会（NLP2004）
【非特許文献４】池田和史，柳原正，松本一則，滝嶋康弘，“くだけた表現を高精度に解析するための正規化ルール自動生成手法”，情報処理学会論文誌，vol3．No.3 pp.68-77，2010
【非特許文献５】Kudo,T.，Japanese Morphological Analyzer,インターネット＜URL:http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html＞

【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかし、非特許文献１及び非特許文献２に記載の教師データを用いた手法により崩れ表記語を抽出する場合、Ｗｅｂデータから、図７のような正解ペアを人手で作成する必要があり、正解ペアの生成コストが高いという課題がある。
【０００６】
また、教師データを用いない手法に基づく場合、獲得候補となる崩れ語の候補が限られた候補（カタカナ語，既存解析器で未知語となった語等）に限られており、多様な崩れ表記を獲得することができないという課題がある。これは、既存解析器では崩れ表記語は誤って解析されてしまうことが多く、多様な崩れ表記語を獲得することが困難なためである。なぜならば、日本語は単語間にスペースなどの区切りが存在しないため、一般に存在するテキストにおいては形態素の正しい区切り位置を解析することが困難である。また、Ｗｅｂ上には、ひらがなや漢字とひらがな、カタカナとひらがな等で書かれる崩れ表記語が多数存在しており、解析が困難である。例えば、「すげー」、「やば」、「さみい」、「サムい」、「寒っ」等である。また、図８に非特許文献５に記載のMecab（IPAdic）を用いて崩れ表記語を含む文を解析した結果の一例を示す。
【０００７】
本発明は、上記問題点を解決するために成されたものであり、効率よく、同義語ペアを獲得することができる同義語ペア獲得装置、方法、及びプログラム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】
【０００８】
上記目的を達成するために、第１の発明に係る同義語ペア獲得装置は、文書から、正規表記語、又は前記正規表記語に対して揺らいだ表記の候補である崩れ表記語である複数の単語分割候補を生成する単語分割候補生成部と、前記単語分割候補生成部により生成された前記複数の単語分割候補に基づいて、前記複数の単語分割候補の各々について、単語の意味ベクトルを計算する意味ベクトル計算部と、正規表記語である前記単語分割候補の各々について、前記意味ベクトルに基づいて計算される意味類似度と、単語の読みに基づいて計算される音類似度とに基づいて、前記単語分割候補を、前記複数の単語分割候補から選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得する同義語ペア獲得部と、を含んで構成されている。
【０００９】
また、第１の発明に係る同義語ペア獲得装置において、前記同義語ペア獲得部は、正規表記語である前記単語分割候補の各々について、前記意味類似度と前記音類似度とに基づいて、前記単語分割候補を、前記複数の単語分割候補から選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得し、選択された前記単語分割候補の各々について、前記意味類似度と前記音類似度とに基づいて、前記単語分割候補を、前記複数の単語分割候補から選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得するようにしてもよい。
【００１０】
第２の発明に係る同義語ペア獲得方法は、単語分割候補生成部が、文書から、正規表記

語、又は前記正規表記語に対して揺らいだ表記の候補である崩れ表記語である複数の単語分割候補を生成するステップと、意味ベクトル計算部が、前記単語分割候補生成部により生成された前記複数の単語分割候補に基づいて、前記複数の単語分割候補の各々について、単語の意味ベクトルを計算するステップと、同義語ペア獲得部が、正規表記語である前記単語分割候補の各々について、前記意味ベクトルに基づいて計算される意味類似度と、単語の読みに基づいて計算される音類似度とに基づいて、前記単語分割候補を、前記複数の単語分割候補から選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得するステップと、を含んで実行することを特徴とする。

【００１１】

また、第２の発明に係る同義語ペア獲得方法は、前記同義語ペア獲得部が獲得するステップは、正規表記語である前記単語分割候補の各々について、前記意味類似度と前記音類似度とに基づいて、前記単語分割候補を、前記複数の単語分割候補から選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得し、選択された前記単語分割候補の各々について、前記意味類似度と前記音類似度とに基づいて、前記単語分割候補を、前記複数の単語分割候補から選択し、正規表記語である前記単語分割候補と、選択された前記単語分割候補とのペアを、同義語ペアとして獲得するようにしてもよい。

【００１２】

第３の発明に係るプログラムは、第１の発明に係る同義語ペア獲得装置の各部として機能させるためのプログラムである。

【発明の効果】

【００１３】

本発明の同義語ペア獲得装置、方法、及びプログラムによれば、文書から、正規表記語、又は崩れ表記語である複数の単語分割候補を生成し、複数の単語分割候補に基づいて、複数の単語分割候補の各々について、単語の意味ベクトルを計算し、正規表記語である単語分割候補の各々について、意味ベクトルに基づいて計算される意味類似度と、単語の読みに基づいて計算される音類似度とに基づいて、単語分割候補を、複数の単語分割候補から選択し、正規表記語である単語分割候補と、選択された単語分割候補とのペアを、同義語ペアとして獲得することにより、効率よく、同義語ペアを獲得することができる、という効果が得られる。

【図面の簡単な説明】

【００１４】

【図１】本発明の実施の形態に係る同義語ペア獲得装置の構成を示すブロック図である。

【図２】音類似度の一例を示す図である。

【図３】同義語ペアの獲得の例を示す概念図である。

【図４】正規表記語を起点として単語分割候補を選択する例を示す図である。

【図５】選択された単語分割候補を起点として更に単語分割候補を選択する例を示す図である。

【図６】本発明の実施の形態に係る同義語ペア獲得装置における同義語ペア獲得処理ルーチンを示すフローチャートである。

【図７】正規表記語及び崩れ表記語の組み合わせの一例を示す図である。

【図８】Ｍｅｃａｂを用いて崩れ表記語を含む文を解析した結果の一例を示す図である。

【発明を実施するための形態】

【００１５】

以下、図面を参照して本発明の実施の形態を詳細に説明する。

【００１６】

＜本発明の実施の形態に係る同義語ペア獲得装置の構成＞

【００１７】

次に、本発明の実施の形態に係る同義語ペア獲得装置の構成について説明する。図１に

示すように、本発明の実施の形態に係る同義語ペア獲得装置１００は、ＣＰＵと、ＲＡＭと、後述する同義語ペア獲得処理ルーチンを実行するためのプログラムや各種データを記憶したＲＯＭと、を含むコンピュータで構成することが出来る。この同義語ペア獲得装置１００は、機能的には図１に示すように入力部１０と、演算部２０と、出力部５０とを備えている。

【００１８】

入力部１０は、崩れ表記語を含む文書からなる文書集合を受け付ける。

【００１９】

演算部２０は、辞書データベース２８と、単語分割候補生成部３０と、意味ベクトル計算部３２と、同義語ペア獲得部３４とを含んで構成されている。

【００２０】

辞書データベース２８には、辞書引きを行うために必要な辞書（読み、表記、品詞）が記憶されている。

【００２１】

単語分割候補生成部３０は、入力部１０により受け付けた文書集合の文書の各々から、正規表記語、又は正規表記語に対して揺らいだ表記の候補である崩れ表記語である複数の単語分割候補を生成する。

【００２２】

単語分割候補生成部３０は、具体的には、文書に対して、既存の単語分割手法である以下の第１の手法から第３の手法の各々を適用して単語分割候補を生成する。この際、辞書データベース２８に存在しない崩れ表記語についても区切り候補として出力できるような手法を用いる。

【００２３】

単語分割候補生成部３０は、文書集合に含まれる文書の各々に対して、第１の手法として、点推定を用いた単語分割手法を適用して単語分割候補の生成を行う。点推定を用いた単語分割手法では、文字ｎｇｒａｍ、文字種ｎｇｒａｍ等を素性とした文字間の区切りモデルを用いて、文書を複数の単語分割候補に分割する。

【００２４】

単語分割候補生成部３０は、文書集合に対して、第２の手法として、教師なし解析を用いた単語分割手法を適用して、単語分割候補の生成を行う。教師なし解析を用いた単語分割手法では、サンプリングした単語分割候補に対して出現頻度等を算出し、目的関数が最適化されるように、文書の各々を単語分割候補に分割する。

【００２５】

単語分割候補生成部３０は、文書集合に含まれる文書の各々に対して、第３の手法として、Ｍｅｃａｂ等による解析結果を取得し、あらかじめ定めたルールを元に一部結合させた単語分割候補の生成を行う。ルールとしては、例えば、未知語連続は結合する、名詞連続は結合する等である。なお、ルールとして以下の方法を用いてもよい。例えば、Ｔｗｉｔｔｅｒ（Ｒ）等から短い文を切り出して、単語分割候補とする場合には、短い文の切り出しは、複数の区切り文字（例えば、改行、記号的表現（「！」，「ｗ」，「♪」）、句読点（「、」，「。」）など）を設定し、短い文を区切り文字で分割するようにすればよい。このように設定することで、例えば「やっべぇぇｗｗｗｗｗｗｗｗｗ」という文であれば、「ｗ」以前の「やっべぇぇ」を単語分割候補として取得できる。また、「おっはよお♪ ってお昼だけど・・・ 今起きた・・・」という文であれば、「♪」以前の「おっはよお」が単語分割候補として取得できる。上記のようにして取得した文字数がｎ文字以下の文字列を形態素辞書に追加して解析を行うようにしてもよい。

【００２６】

意味ベクトル計算部３２は、単語分割候補生成部３０により生成された複数の単語分割候補に基づいて、複数の単語分割候補の各々について、単語の意味ベクトルを計算する。

【００２７】

意味ベクトル計算部３２は、具体的には、単語分割候補生成部３０により生成された複

数の単語分割候補を列挙するように、単語区切りが付与された文書集合に対し、単語分割候補として出現した各単語の意味ベクトルを計算する。この際、各単語の意味ベクトルを求める手法としては既存の手法を用いることができる。例えば、非特許文献６に記載のｗｏｒｄ２ｖｅｃ等が代表的な手法として挙げられる。

【００２８】

［非特許文献６］：Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

【００２９】

同義語ペア獲得部３４は、正規表記語である単語分割候補の各々について、意味ベクトルに基づいて計算される意味類似度が閾値以上であって、かつ、単語の読みに基づいて計算される音類似度が閾値以上となる、単語分割候補を、複数の単語分割候補から選択し、正規表記語である単語分割候補と、選択された単語分割候補とのペアを、同義語ペアとして獲得する。同義語ペア獲得部３４は、更に、選択された単語分割候補の各々について、意味ベクトルに基づいて計算される意味類似度が閾値以上であって、かつ、単語の読みに基づいて計算される音類似度が閾値以上となる、単語分割候補を、複数の単語分割候補から選択し、正規表記語である単語分割候補と、選択された単語分割候補とのペアを、同義語ペアとして獲得する。図３に同義語ペア獲得部３４の処理の概念図を示す。

【００３０】

同義語ペア獲得部３４は、具体的には、まず正規表記語である単語分割候補の各々について、他の単語分割候補の各々との意味類似度の計算を行う。意味類似度は、意味ベクトル計算部３２において求めた単語ごとの意味ベクトルのコサイン類似度を用いて計算する。

【００３１】

同義語ペア獲得部３４は、次に、正規表記語である単語分割候補の各々について、他の単語分割候補との音類似度の計算を行う。本実施の形態では、音類似度として、音類似度距離を、単語分割候補の読みに基づいて計算する。ここで、漢字表記は読み推定を行い、カタカナ表記はひらがなに変換する。変換コストは次のように設定する。同一文字の変換コストは０とする。また、母音(小文字も含む（例：ぁ，ぃ，ぅ，ぇ，ぉ）)、促音（っ）、撥音（ん）、長音の削除はコスト０とする。ただし、単語の先頭における削除はコスト１として音類似度距離をカウントアップする。また、同行又は同列（日本語ひらがな50音表の同行又は同列を指す。濁音又は半濁音は濁音化又は半濁音化する前の文字と同一の位置として考える）文字の置換、母音-促音間の置換、母音‐長音間、母音‐母音間の変換はコスト０とする。例えば、「ぶ」又は「ぷ」→「ふ」というような同行又は同列の文字列（はひふへほうくすつぬむゆる）をコスト０とする。上記以外の変換はコスト１として音類似度距離をカウントアップする。図２に音類似度距離の計算例を示す。本実施の形態では、閾値以上の音類似度のものをフィルタリングするため、音類似度距離が閾値以下のものがフィルタリングされる。

【００３２】

次に、同義語ペア獲得部３４は、文書集合から得られた正規表記語の単語分割候補の各々について、以下に説明する第１の獲得処理及び第２の獲得処理を行って、同義語ペアを獲得する。同義語ペア獲得部３４の第１の獲得処理では、文書集合から得られた正規表記語の単語分割候補の各々について、以下の処理を行う。

【００３３】

まず、当該正規表記語の単語分割候補について、文書集合中に現れた他の単語分割候補から、他の単語分割候補との間の意味類似度が予め定めた閾値以上である単語分割候補をフィルタリングする。次に、フィルタリングされた単語分割候補から、当該正規表記語について、他の単語分割候補との音類似度が予め定めた閾値以上（音類似度距離が閾値以下）となる単語分割候補をフィルタリングする。更に、フィルタリングされた単語分割候補から、辞書データベース２８において、当該単語分割候補の表記が辞書中の正規表記語と

して存在し、かつ辞書中の当該正規表記語の品詞と同一の品詞であるものを削除する。そして、同義語ペア獲得部３４は、削除後の単語分割候補を選択する。このようにして、当該正規表記語の単語分割候補と選択した単語分割候補とのペアを、同義語ペアとして獲得とする。図４に第１の獲得処理の一例を示す。図４では、正規表記語の単語分割候補「さむい」を起点として単語分割候補を選択している。

【００３４】

次に、同義語ペア獲得部３４は、当該正規表記語の単語分割候補について、以下のように、上記の第１の獲得処理で当該正規表記語の単語分割候補について同義語ペアとして選択された単語分割候補を起点とした、第２の獲得処理を行う。まず、上記の第１の獲得処理で当該正規表記語の単語分割候補について同義語ペアとして選択された単語分割候補の各々について、他の単語分割候補との間の意味類似度の計算、及び音類似度距離の計算を行う。次に、当該正規表記語の単語分割候補について同義語ペアとして選択された単語分割候補の各々について、以下の処理を行う。

【００３５】

当該単語分割候補について、文書集合中に現れた他の単語分割候補の各々との間の意味類似度が予め定めた閾値以上である単語分割候補をフィルタリングする。次に、フィルタリングされた単語分割候補から、当該単語分割候補との音類似度距離が予め定めた閾値以下となる単語分割候補をフィルタリングする。更に、フィルタリングされた単語分割候補から、辞書データベース２８において、単語分割候補の表記が辞書中の正規表記語として存在し、かつ辞書中の当該正規表記語の品詞と同一の品詞であるものを削除する。そして、同義語ペア獲得部３４は、削除後の単語分割候補を選択する。このようにして、当該正規表記語の単語分割候補と選択した単語分割候補とのペアを、同義語ペアとして獲得とする。図５に第２の獲得処理の一例を示す。図５では、第１の獲得処理で正規表記語の単語分割候補「さむい」に対して選択された単語分割候補「さみぃ」を起点として単語分割候補を選択している。更に、同義語ペア獲得部３４は、上記第２の獲得処理で選択された単語分割候補を起点として、上記第２の獲得処理と同じ処理を予め定めた回数繰り返し、当該正規表記語の単語分割候補と選択した単語分割候補とのペアを、同義語ペアとして獲得する。

【００３６】

＜本発明の実施の形態に係る同義語ペア獲得装置の作用＞

【００３７】

次に、本発明の実施の形態に係る同義語ペア獲得装置１００の作用について説明する。入力部１０において崩れ表記語を含む文書からなる文書集合を受け付けると、同義語ペア獲得装置１００は、図６に示す同義語ペア獲得処理ルーチンを実行する。

【００３８】

まず、ステップＳ１００では、入力部１０において受け付けた文書集合の文書の各々から複数の単語分割候補を生成する。

【００３９】

次に、ステップＳ１０２では、ステップＳ１００で生成された複数の単語分割候補に基づいて、単語分割候補の各々について、単語の意味ベクトルを計算する。

【００４０】

ステップＳ１０４では、ステップＳ１００で生成された正規表記語である単語分割候補の各々について、ステップＳ１０２で計算された意味ベクトルに基づいて、他の単語分割候補の各々との意味類似度を計算する。

【００４１】

ステップＳ１０６では、ステップＳ１００で生成された正規表記語である単語分割候補の各々について、単語分割候補の読みに基づいて他の単語分割候補の各々との音類似度距離を計算する。

【００４２】

ステップＳ１０８では、正規表記語である単語分割候補の各々について、ステップＳ１

０４で計算された意味類似度が閾値以上であって、かつ、ステップＳ１０６で計算された音類似度距離が閾値以下となる、単語分割候補を、複数の単語分割候補から選択し、正規表記語である単語分割候補と、選択された単語分割候補とのペアを、同義語ペアとして獲得する。

【００４３】

ステップＳ１１０では、正規表記語である単語分割候補の各々に対し、ステップＳ１０８又は前回のステップＳ１１０で選択された単語分割候補の各々について、ステップＳ１０４と同様に計算される意味類似度が閾値以上であって、かつ、ステップＳ１０６と同様に計算される音類似度距離が閾値以下となる、単語分割候補を、複数の単語分割候補から選択し、当該正規表記語である単語分割候補と、選択された単語分割候補とのペアを、同義語ペアとして獲得する。

【００４４】

ステップＳ１１２では、ステップＳ１１０の処理を予め定めた回数繰り返したかを判定し、繰り返していればステップＳ１１４へ移行し、繰り返していなければステップＳ１１０へ戻って処理を繰り返す。

【００４５】

ステップＳ１１４では、ステップＳ１０８及びステップＳ１１０で獲得された同義語ペアを出力部５０に出力して処理を終了する。

【００４６】

以上説明したように、本発明の実施の形態に係る同義語ペア獲得装置によれば、文書から、正規表記語、又は崩れ表記語である複数の単語分割候補を生成し、複数の単語分割候補に基づいて、複数の単語分割候補の各々について、単語の意味ベクトルを計算し、正規表記語である単語分割候補の各々について、意味ベクトルに基づいて計算される意味類似度が閾値以上であって、かつ、単語の読みに基づいて計算される音類似度距離が閾値以下となる、単語分割候補を、複数の単語分割候補から選択し、正規表記語である単語分割候補と、選択された単語分割候補とのペアを、同義語ペアとして獲得することにより、効率よく、同義語ペアを獲得することができる。

【００４７】

また、意味類似度と音類似度の双方を考慮することにより、精度よく同義候補のペアを獲得することができる。

【００４８】

また、正規表記語を起点とした獲得だけではフィルタされてしまった単語分割候補に対しても、選択された単語分割候補を起点として新たな同義語ペアを獲得することでより多様な崩れ表記語を獲得することが可能になる。

【００４９】

また、従来手法に比べ、多様な崩れ表記語の正しい区切りとして単語分割候補を生成することが可能になる。

【００５０】

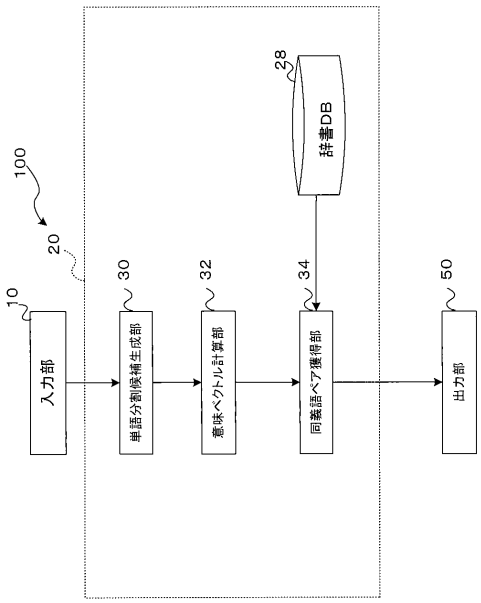
なお、本発明は、上述した実施の形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

【符号の説明】

【００５１】

１０　入力部
２０　演算部
２８　辞書データベース
３０　単語分割候補生成部
３２　意味ベクトル計算部
３４　同義語ペア獲得部
５０　出力部
１００　同義語ペア獲得装置

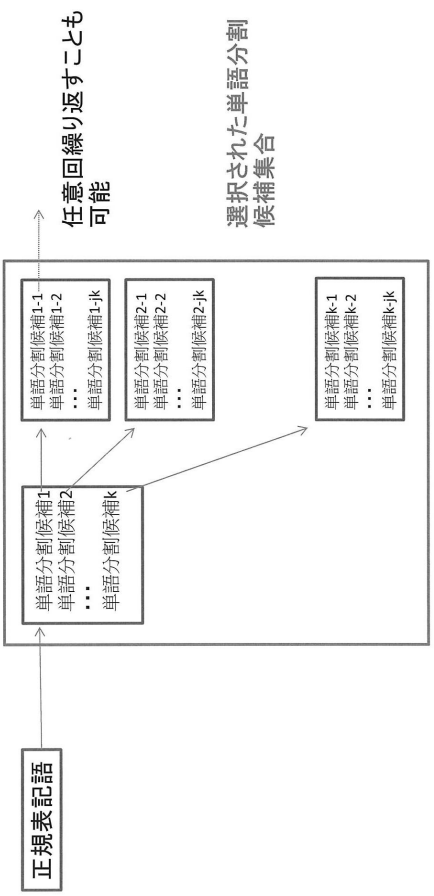
【図１】



【図２】

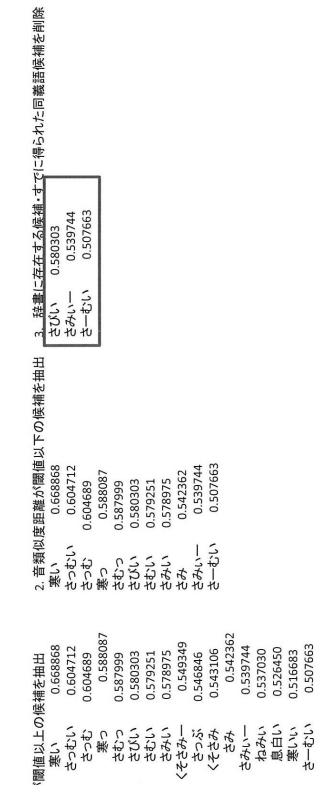
単語1	単語2	文字アライメント	音類似度距離	備考
さむい	さみい	さ-さ/む-み/い-い	0	「む-み」の変換は同列置換なのでコスト0
さむい	さみい	さ-さ/む-み/い-い	0	「い-い」の変換は母音置換なのでコスト0
さむい	さぶい	さ-さ/む-ぶ/い-い	0	「む-ぶ」の変換は同行置換なのでコスト0
さむい	ねむい	さ-ね/む-む/い-い	1	「さ-ね」の変換はその他置換なのでコスト1

【図３】

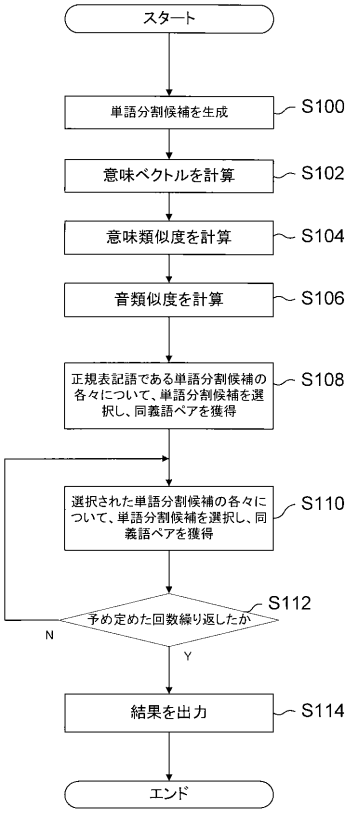


【図４】

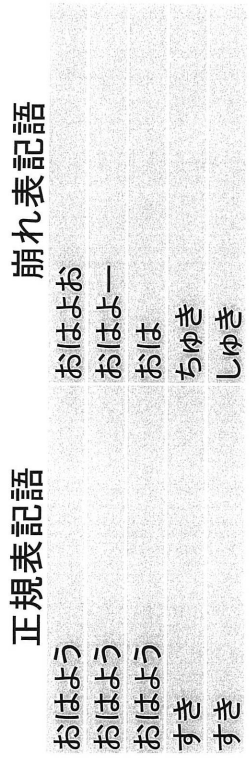
【図５】


1. 意味類似度が閾値以上の候補を抽出
寒い 0.668868
さっむい 0.604712
さっむ 0.604689
寒っ 0.588087
さむっ 0.587999
さびい 0.580303
さむい 0.579251
さみい 0.578975
くそさみー 0.549349
さっぶ 0.546846
くそさみ 0.543106
さみ 0.542362
さみいー 0.539744
ねみい 0.537030
息白い 0.526450
寒しい 0.516683
さーむい 0.507663
2. 音類似度語尾が閾値以下の候補を抽出
寒い 0.668868
さっむい 0.604712
さっむ 0.604689
寒っ 0.588087
さむっ 0.587999
さびい 0.580303
さむい 0.579251
さみい 0.578975
さみ 0.542362
さみいー 0.539744
さーむい 0.507663
3. 話者に在在する候補・すでに得られた同義語候補を削除
さびい 0.580303
さみいー 0.539744
さーむい 0.507663

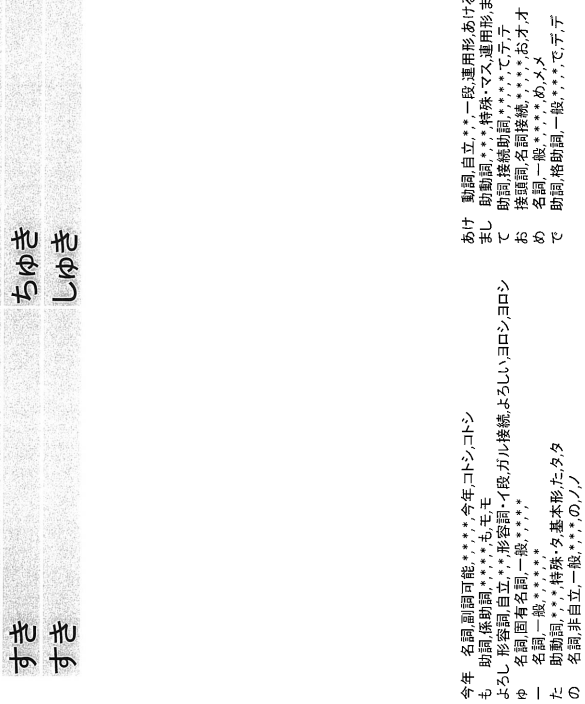
【図６】


スタート
単語分割候補を生成
S100
意味ベクトルを計算
S102
意味類似度を計算
S104
音類似度を計算
S106
正規表記語である単語分割候補の各々について、単語分割候補を選択し、同義語ペアを獲得
S108
選択された単語分割候補の各々について、単語分割候補を選択し、同義語ペアを獲得
S110
予め定めた回数繰り返したか
S112
N
Y
結果を出力
S114
エンド

【図７】


正規表記語
崩れ表記語
おはよう
おはよお
おはよう
おはよー
おはよう
おは
すき
ちゆき
すき
しゆき

【図８】


今年 名詞,副詞可能,*,*,*,今年,コトシ,コトシ
も 助詞,係助詞,*,*,*,*,も,モ,モ
よろし 形容詞,自立,*,*,形容詞・イ段,ガル接続,よろしい,ヨロシ,ヨロシ
ゆ 名詞,固有名詞,一般,*,*,*
ー 名詞,一般,*,*,*,*
た 助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
の 名詞,非自立,一般,*,*,*,の,ノ,ノ
も 助詞,係助詞,*,*,*,*,も,モ,モ
あけ 動詞,自立,*,*,一段,連用形,あける,アケ,アケ
まし 助動詞,*,*,*,特殊・マス,連用形,ます,マシ,マシ
て 助詞,接続助詞,*,*,*,*,て,テ,テ
お 接頭詞,名詞接続,*,*,*,*,お,オ,オ
め 名詞,一般,*,*,*,*,め,メ,メ
で 助詞,格助詞,一般,*,*,*,で,デ,デ

フロントページの続き

(72)発明者　松尾　義博
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

審査官　松尾　真人

(56)参考文献　特開２０００－２２２４２７（ＪＰ，Ａ）
特開２００９－１７６１４８（ＪＰ，Ａ）
特開２００８－２９９８６８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２７－１７／３０