

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号
特許第6368633号
(P6368633)

(45) 発行日 平成30年8月1日(2018.8.1)

(24) 登録日 平成30年7月13日(2018.7.13)

(51) Int.Cl.

F I

G O 6 F 17/27 (2006.01)

G O 6 F 17/30 (2006.01)

G O 6 F 17/27 6 9 5

G O 6 F 17/30 3 2 O D

G O 6 F 17/30 3 5 O C

G O 6 F 17/30 1 7 O A

請求項の数 8 (全 18 頁)

(21) 出願番号	特願2014-248130 (P2014-248130)	(73) 特許権者	000004226
(22) 出願日	平成26年12月8日 (2014.12.8)		日本電信電話株式会社
(65) 公開番号	特開2016-110440 (P2016-110440A)		東京都千代田区大手町一丁目5番1号
(43) 公開日	平成28年6月20日 (2016.6.20)	(74) 代理人	110001519
審査請求日	平成28年12月9日 (2016.12.9)		特許業務法人太陽国際特許事務所
		(72) 発明者	藤野 昭典
			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	鈴木 潤
			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	平尾 努
			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】用語意味学習装置、用語意味判定装置、方法、及びプログラム

(57) 【特許請求の範囲】

【請求項1】

入力された、用語の意味コードが予め付与された複数の用語の各々について、前記用語の文字列に含まれる部分文字列から、各部分文字列の有無又は個数を表す特徴ベクトルを生成する特徴ベクトル生成部と、

前記特徴ベクトル生成部により前記複数の用語の各々について生成された前記特徴ベクトルと、前記複数の用語の各々に付与された意味コードとに基づいて、用語の意味コードを判定するための判定ルールを生成する判定ルール生成部と、

を含む用語意味学習装置であって、

前記用語について生成された前記特徴ベクトルが表す部分文字列に含まれる全ての文字を要素とする集合と、前記用語の文字列に含まれる全ての文字を要素とする集合とは一致する、用語意味学習装置。

【請求項2】

前記複数の用語の各々は、用語、及び意味コードの対の集合である訓練データとして入力され、

前記訓練データを構成する各用語は、特定の専門分野において唯一の意味が定義されている用語とし、

前記判定ルール生成部は、前記訓練データにおいて意味コードkが付与された用語を正例、前記意味コードk以外の意味コードが付与された用語を負例として、前記判定ルールを生成する、

10

20

請求項 1 に記載の用語意味学習装置。

【請求項 3】

入力された用語の文字列に含まれる部分文字列から、各部分文字列の有無又は個数を表す特徴ベクトルを生成する特徴ベクトル生成部と、

前記特徴ベクトル生成部により生成された前記特徴ベクトルと、あらかじめ学習された、用語の意味コードを判定するための判定ルールとに基づいて、前記入力された用語の意味コードを判定する意味コード判定部と、

を含む用語意味判定装置であって、

前記入力された用語について生成された前記特徴ベクトルが表す部分文字列に含まれる全ての文字を要素とする集合と、前記入力された用語の文字列に含まれる全ての文字を要素とする集合とは一致する、用語意味判定装置。

10

【請求項 4】

前記判定ルールは、特定の専門分野において唯一の意味が定義されている用語、及び意味コードの対の集合である訓練データを用いて、前記入力された用語の特徴量ベクトルを入力として、前記入力された用語の意味コードを判定するよう、あらかじめ学習されたものとする、

請求項 3 に記載の用語意味判定装置。

【請求項 5】

前記特徴ベクトル生成部は、前記入力された用語の文字の文字種を用いて定義される特徴量を追加した前記特徴ベクトルを生成する請求項 3 又は請求項 4 に記載の用語意味判定装置。

20

【請求項 6】

特徴ベクトル生成部が、入力された、用語の意味コードが予め付与された複数の用語の各々について、前記用語の文字列に含まれる部分文字列から、各部分文字列の有無又は個数を表す特徴ベクトルを生成するステップと、

判定ルール生成部が、前記特徴ベクトル生成部により前記複数の用語の各々について生成された前記特徴ベクトルと、前記複数の用語の各々に付与された意味コードとに基づいて、用語の意味コードを判定するための判定ルールを生成するステップと、

を含む用語意味学習方法であって、

前記用語について生成された前記特徴ベクトルが表す部分文字列に含まれる全ての文字を要素とする集合と、前記用語の文字列に含まれる全ての文字を要素とする集合とは一致する、用語意味学習方法。

30

【請求項 7】

コンピュータを、請求項 1 又は請求項 2 に記載の用語意味学習装置を構成する各部として機能させるためのプログラム。

【請求項 8】

コンピュータを、請求項 3 ～ 5 の何れか 1 項に記載の用語意味判定装置を構成する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】

40

【0001】

本発明は、用語意味学習装置、用語意味判定装置、方法、及びプログラムに係り、特に、用語に対する意味コードを判定するための用語意味学習装置、意味判定装置、方法、及びプログラムに関する。

【背景技術】

【0002】

用語には、ひらがな、カタカナ、及び漢字などの異なる文字を用いた表記や、略語や省略などのような表記揺れがある。また、同義語及び類義語などのように同じ意味を表すが異なる表現の用語がある。このような異なる表記又は表現の用語に同じ意味であることを表す意味コードを付与する問題に対しては、非特許文献 1 に記載されている、用語のペア

50

が同義語であるかどうかを文脈類似度と表記類似度を用いて推定する技術を利用できる。また、非特許文献2に記載されている、「スパゲッティ」と「スパゲッティー」など、一部の表記が異なる編集距離の小さい異表記対を抽出する技術を利用できる。あるいは、非特許文献3に記載されている、実際の文書と一緒に用いられる単語の情報をもとに用語を分類する技術を利用できる。

【先行技術文献】

【非特許文献】

【0003】

【非特許文献1】森本康嗣，柳井孝介，岩山真，「文脈類似度と表記類似度を用いた教師あり同義語抽出」，言語処理学会第16回年次大会発表論文集，974 - 977 (2010)。

10

【非特許文献2】小島正裕，村田真樹，風間淳一，黒田航，藤田篤，荒牧英治，土田正明，渡辺靖彦，鳥澤健太郎，「機械学習と種々の素性を用いた編集距離の小さい日本語異表記対の抽出」，言語処理学会第16回年次大会発表論文集，928 - 931 (2010)。

【非特許文献3】荒牧英治，今井健，梶野正幸，美代賢吾，大江和彦，「医学辞書を用いた用語間関係の自動抽出手法と用語の自動分類手法に関する研究」，医療情報学，Vol.25，No.6, pp.463-474，2006。

【発明の概要】

【発明が解決しようとする課題】

【0004】

本発明が解決しようとする課題は、事前に人手で定義されている意味コードを専門用語に付与する問題において、非常に多数の意味コードの候補の中から、各用語に最も適した意味コードを1つ選択することである。

20

【0005】

一般の単語と異なり、正確に内容を伝達するために、大部分の専門用語には専門分野ごとにそれぞれ唯一の意味が定義されている、と考えられる。したがって、専門分野が既知の文書に含まれる専門用語に意味コードを付与する問題は、専門用語に唯一の意味が定義されているという仮定のもとで、同じ意味をもつ異なる表記又は表現で書かれた用語に同じ意味コードを付与することに相当する。

【0006】

従来の技術では、非特許文献1、及び非特許文献2に記載のように、2つの用語が同じ意味をもつかどうかを推定するか、非特許文献3のように、クラスタリング技術を利用して用語の集合を類似の意味をもつ用語に分けることが行われている。本発明が解決しようとする課題に非特許文献1、及び非特許文献2の技術を適用する場合、例えば、意味コードごとに用語例を1つ選択し、意味コードを付与したい用語がどの用語例に近いかを該当技術で調べることで意味コードを推定できる。

30

【0007】

しかし、この方法では、用語例の選択の仕方に意味コードの推定精度が大きく影響される可能性がある。一方、本発明が解決しようとする課題に非特許文献3の技術を用いる場合、クラスタリング技術で機械的に分類された用語の集合と、人手で定義された意味コードとを一対一に対応付けることは容易ではない。専門用語に対して付与すべき意味コードを推定する問題に対して、高い推定精度を達成する手法を開発することが課題である。

40

【0008】

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、精度よく用語の意味コードを判定するための用語意味学習装置、用語意味判定装置、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】

【0009】

上記目的を達成するために、第1の発明に係る用語意味学習装置は、入力された、用語の意味コードが予め付与された複数の用語の各々について、前記用語の文字列に含まれる部分文字列から、各部分文字列の有無又は個数を表す特徴ベクトルを生成する特徴ベクトル

50

ル生成部と、前記特徴ベクトル生成部により前記複数の用語の各々について生成された前記特徴ベクトルと、前記複数の用語の各々に付与された意味コードとに基づいて、用語の意味コードを判定するための判定ルールを生成する判定ルール生成部と、を含んで構成されている。

【0010】

また、第1の発明に係る用語意味学習装置において、前記用語を、特定の専門分野において唯一の意味が定義されている専門用語としてもよい。

【0011】

また、第1の発明に係る用語意味学習装置において、前記特徴ベクトル生成部は、前記複数の用語の各々について、前記用語に含まれる文字、及び前記用語の文字列に含まれる部分文字列から、各文字の有無又は個数と、各部分文字列の有無又は個数とを表す前記特徴ベクトルを生成してもよい。

10

【0012】

また、第1の発明に係る用語意味学習装置において、前記特徴ベクトル生成部は、前記複数の用語の各々について、前記用語に含まれる文字、及び前記用語の文字列に含まれる部分文字列から、各文字の有無又は個数と、各部分文字列の有無又は個数と、前記用語に含まれる各文字の文字種とを表す前記特徴ベクトルを生成してもよい。更に、前記用語にふうまれる各部分文字列の文字種を更に表す前記特徴ベクトルを生成してもよい。

【0013】

また、第1の発明に係る用語意味学習装置において、入力された、用語の意味コードが予め付与された複数の用語と、同義語又は類義語が格納されている辞書とに基づいて、前記意味コードが予め付与されている用語の同義語又は類似語となる用語に、前記意味コードを付与する同義語類義語意味コード付与部を更に含み、前記特徴ベクトル生成部は、前記用語の意味コードが予め付与された複数の用語、及び同義語類義語意味コード付与部によって意味コードが付与された前記同義語又は類似語となる用語の各々について、前記特徴ベクトルを生成するように構成してもよい。

20

【0014】

また、上記判定ルール生成部は、前記特徴ベクトル生成部により前記複数の用語の各々について生成された前記特徴ベクトルと、前記複数の用語の各々に付与された意味コードとに基づいて、前記意味コード毎に、用語に前記意味コードを付与するか否かを識別するための2値分類器を、前記判定ルールとして生成するようにしてもよい。

30

【0015】

第2の発明に係る用語意味判定装置は、入力された用語の文字列に含まれる部分文字列から、各部分文字列の有無又は個数とを表す特徴ベクトルを生成する特徴ベクトル生成部と、前記特徴ベクトル生成部により生成された前記特徴ベクトルと、あらかじめ学習された、用語の意味コードを判定するための判定ルールとに基づいて、前記用語の意味コードを判定する意味コード判定部と、を含んで構成されている。

【0016】

また、第2の発明に係る用語意味判定装置において、前記用語を、特定の専門分野において唯一の意味が定義されている専門用語としてもよい。

40

【0017】

また、第2の発明に係る用語意味判定装置において、前記特徴ベクトル生成部は、前記用語について、前記用語に含まれる文字、及び前記用語の文字列に含まれる部分文字列から、各文字の有無又は個数と、各部分文字列の有無又は個数とを表す前記特徴ベクトルを生成してもよい。

【0018】

第1の発明に係る用語意味学習方法は、特徴ベクトル生成部が、入力された、用語の意味コードが予め付与された複数の用語の各々について、前記用語の文字列に含まれる部分文字列から、各部分文字列の有無又は個数とを表す特徴ベクトルを生成するステップと、判定ルール生成部が、前記特徴ベクトル生成部により前記複数の用語の各々について生成さ

50

れた前記特徴ベクトルと、前記複数の用語の各々に付与された意味コードとに基づいて、用語の意味コードを判定するための判定ルールを生成するステップと、を含んで実行することを特徴とする。

【 0 0 1 9 】

第 3 の発明に係るプログラムは、コンピュータを、上記第 1 の発明に係る用語意味学習装置、又は上記第 2 の発明に係る用語意味判定装置を構成する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】

【 0 0 2 0 】

本発明の用語意味学習装置、用語意味判定装置、方法、及びプログラムによれば、複数の用語の各々について、用語に含まれる文字又は部分文字列から、特徴ベクトルを生成し、複数の用語の各々について生成された特徴ベクトルと、複数の用語の各々に付与された意味コードとに基づいて、意味コードを判定するための判定ルールを生成することで、精度よく用語の意味コードを判定するための判定ルールを学習することができる、という効果が得られる。

【 0 0 2 1 】

また、用語意味判定装置、方法、及びプログラムによれば、用語に含まれる文字又は部分文字列から、特徴ベクトルを生成し、生成された特徴ベクトルと、予め学習された判定ルールとに基づいて、用語の意味コードを判定することで、精度よく用語の意味コードを判定することができる、という効果が得られる。

【図面の簡単な説明】

【 0 0 2 2 】

【図 1】本発明の第 1 ～ 第 3 の実施の形態に係る用語意味学習装置の機能的構成を示すブロック図である。

【図 2】専門用語に含まれる部分文字列に対応する特徴量のインデックスを示す概念図である。

【図 3】本発明の実施の形態に係る用語意味判定装置の機能的構成を示すブロック図である。

【図 4】本発明の第 1 ～ 第 3 の実施の形態に係る用語意味学習装置における用語意味学習処理ルーチンのフローチャート図である。

【図 5】本発明の実施の形態に係る用語意味判定装置における用語意味判定処理ルーチンのフローチャート図である。

【図 6】専門用語に含まれる文字と、部分文字列とに対応する特徴量のインデックスを示す概念図である。

【図 7】専門用語に含まれる文字と、部分文字列と、各文字の文字種とに対応する特徴量のインデックスを示す概念図である。

【図 8】本発明の第 4 の実施の形態に係る用語意味学習装置の機能的構成を示すブロック図である。

【図 9】本発明の第 4 の実施の形態に係る用語意味学習装置における用語意味学習処理ルーチンのフローチャート図である。

【発明を実施するための形態】

【 0 0 2 3 】

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。

【 0 0 2 4 】

< 本発明の第 1 の実施の形態に係る用語意味学習装置の構成 >

【 0 0 2 5 】

まず、本発明の第 1 の実施の形態に係る用語意味学習装置の構成について説明する。

【 0 0 2 6 】

図 1 に示すように、本発明の第 1 の実施の形態に係る用語意味学習装置 100 は、CPU と、RAM と、後述する用語意味学習処理ルーチンを実行するためのプログラムや各種

10

20

30

40

50

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。この用語意味学習装置100は、機能的には図1に示すように入力部10と、演算部20とを備えている。

【0027】

入力部10は、専門用語の意味コードが予め付与された複数の専門用語を訓練データ集合として受け付け、訓練データ集合データベース28に格納する。入力される、意味コードが予め付与された複数の専門用語は、意味コードを定義するのに用いられている専門用語、人手で意味コードを付与した実際の文書から抽出した専門用語、及び同義語及び類義語の関係が記載されている任意の辞書から抽出した専門用語を用いて構成される。任意の辞書に含まれている専門用語には、通常、意味コードが付与されていないので、辞書から抽出する際に、意味コードを定義するのに用いられている専門用語と、辞書に記載の同義語及び類義語の関係と、を利用して意味コードを付与する。例えば、意味コードを定義するのに用いられている専門用語と同義語及び類義語の関係にある専門用語に対して同一の意味コードを付与すればよい。

10

【0028】

演算部20は、訓練データ集合データベース28と、特徴ベクトル生成部30と、判定ルール生成部32と、判定ルール40とを含んで構成されている。

【0029】

特徴ベクトル生成部30は、訓練データ集合データベース28に格納されている複数の専門用語の各々について、当該専門用語の文字列に含まれる部分文字列から、各部分文字列の個数を表す特徴ベクトルを生成する。なお、各部分文字列の有無を表す特徴ベクトルを生成してもよい。

20

【0030】

特徴ベクトル生成部30では、専門用語を構成する部分文字列、例えば、文字ベースのbigram、又はtrigramなどを特徴量とし、専門用語に含まれる各特徴量の個数を表す特徴ベクトルを生成する。例えば、図2に示すような、「ピロリ菌」という専門用語の文字列に含まれる部分文字列に対応する特徴量のインデックスを仮定する。このとき、専門用語「ピロリ菌」は、 $i \in \{b_1, b_2, b_3, b_4, b_5, t_1, t_2, t_3, t_4\}$ のときに $x_i = 1$ 、 i が、「ピロリ菌」に含まれない部分文字列の特徴量のときに $x_i = 0$ の要素からなる特徴ベクトル $x = (x_1, \dots, x_i, \dots, x_I)$ で表すことができる。なお、ベクトルを表す記号には、記号の前に「 $\wedge$ 」を付して表現する。また、図2において、「 $_$ 」が記号の先頭に付いているものは専門用語の文字列の先端を表し、記号の末尾に付いているものは専門用語の文字列の終端を表す。また、例えば、ひらがな、カタカナ、又はアルファベットなど、専門用語に含まれる文字の文字種を用いて定義される特徴量を追加で用いてもよい。

30

【0031】

判定ルール生成部32は、特徴ベクトル生成部30により複数の専門用語の各々について生成された特徴ベクトルと、複数の専門用語の各々に付与された意味コードとに基づいて、用語の意味コードを判定するための判定ルールを生成する。

【0032】

判定ルール生成部32は、具体的には、複数の専門用語の各々について生成された特徴ベクトルと、複数の専門用語の各々に付与された意味コードとに基づいて、例えば、サポートベクトルマシン(SVM)、あるいはロジスティック回帰モデルのような2値分類器を利用して、意味コードk毎に、スコア関数のパラメータ $\wedge w_k$ の値を推定することにより、判定ルールを機械的に生成し、判定ルール40として記憶する。本実施の形態では、意味コードkのスコア関数のパラメータ $\wedge w_k$ に対して、意味コードkが付与された専門用語を正例、k以外の意味コードが付与された専門用語を負例として、2値分類器の目的関数を最大化させるパラメータ $\wedge w_k$ を計算する。

40

【0033】

< 本発明の第1の実施の形態に係る用語意味判定装置の構成 >

50

【 0 0 3 4 】

次に、本発明の第 1 の実施の形態に係る用語意味判定装置の構成について説明する。図 3 に示すように、本発明の第 1 の実施の形態に係る用語意味判定装置 2 0 0 は、C P U と、R A M と、後述する用語意味判定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 R O M と、を含むコンピュータで構成することが出来る。この用語意味判定装置 2 0 0 は、機能的には図 3 に示すように入力部 2 1 0 と、演算部 2 2 0 と、出力部 2 5 0 とを備えている。

【 0 0 3 5 】

入力部 2 1 0 は、用語の入力を受け付ける。

【 0 0 3 6 】

演算部 2 2 0 は、特徴ベクトル生成部 2 3 0 と、意味コード判定部 2 3 2 と、判定ルール 2 4 0 とを含んで構成されている。

【 0 0 3 7 】

判定ルール 2 4 0 には、判定ルール 4 0 と同じ判定ルールが記憶されている。

【 0 0 3 8 】

特徴ベクトル生成部 2 3 0 は、入力部 2 1 0 で受け付けた用語の文字列に含まれる部分文字列から、特徴ベクトル生成部 3 0 と同様に、各部分文字列の個数を表す特徴ベクトルを生成する。なお、各部分文字列の有無を表す特徴ベクトルを生成してもよい。

【 0 0 3 9 】

意味コード判定部 2 3 2 は、特徴ベクトル生成部 2 3 0 により生成された特徴ベクトルと、判定ルール 2 4 0 とに基づいて、入力部 2 1 0 で受け付けた用語の意味コードを判定する。

【 0 0 4 0 】

意味コード判定部 2 3 2 は、具体的には、特徴ベクトル生成部 2 3 0 で生成した特徴ベクトル $\hat{x}$ と、判定ルール 2 4 0 とを用いて、意味コード k 毎に、判定ルール 2 4 0 のスコア関数のパラメータ $\hat{w}_k$ の値に従って、特徴ベクトル $\hat{x}$ で表される用語が意味コード k に関連する度合いを表すスコア関数 $f_k(\hat{x}; \hat{w}_k)$ を計算する。そして、スコア関数 $f_k(\hat{x}_m; \hat{w}_k)$ の値が最大になる意味コード k を、当該用語に適した意味コードとして選択することで、当該用語の意味コードを判定し、出力部 2 5 0 により出力する。このように予め機械的に学習しておいた判定ルール 2 4 0 を用いて用語の意味コードを判定することで、事前に定義されている意味コードの中から最も適した意味コードを 1 つ選択することができる。

【 0 0 4 1 】

< 本発明の第 1 の実施の形態に係る用語意味学習装置の作用 >

【 0 0 4 2 】

次に、本発明の第 1 の実施の形態に係る用語意味学習装置 1 0 0 の作用について説明する。入力部 1 0 において、専門用語の意味コードが予め付与された複数の専門用語を訓練データ集合として受け付け、訓練データ集合データベース 2 8 に格納すると、用語意味学習装置 1 0 0 は、図 4 に示す用語意味学習処理ルーチンを実行する。

【 0 0 4 3 】

まず、ステップ S 1 0 0 では、訓練データ集合データベース 2 8 に格納されている複数の専門用語の各々を取得する。

【 0 0 4 4 】

次に、ステップ S 1 0 2 では、ステップ S 1 0 0 で取得した複数の専門用語の各々について、当該専門用語の文字列に含まれる部分文字列から、各部分文字列の個数を表す特徴ベクトルを生成する。

【 0 0 4 5 】

ステップ S 1 0 4 では、ステップ S 1 0 2 で複数の専門用語の各々について生成された特徴ベクトルと、ステップ S 1 0 0 で取得した複数の専門用語の各々に付与された意味コードとに基づいて、意味コード k 毎に、スコア関数のパラメータ $\hat{w}_k$ の値を推定するこ

10

20

30

40

50

とにより、判定ルールを生成し、判定ルール 40 に記憶して、処理を終了する。

【0046】

以上説明したように、第 1 の実施の形態に係る用語意味学習装置によれば、複数の専門用語の各々について、用語に含まれる部分文字列から、特徴ベクトルを生成し、複数の専門用語の各々について生成された特徴ベクトルと、複数の専門用語の各々に付与された意味コードとに基づいて、意味コードを判定するための判定ルールを生成することで、精度よく専門用語の意味コードを判定するための判定ルールを学習することができる。

【0047】

< 本発明の第 1 の実施の形態に係る用語意味判定装置の作用 >

【0048】

次に、本発明の第 1 の実施の形態に係る用語意味判定装置 200 の作用について説明する。入力部 210 において用語を受け付けると、用語意味判定装置 200 は、図 5 に示す用語意味判定処理ルーチンを実行する。

【0049】

まず、ステップ S200 では、入力部 10 において受け付けた用語を取得する。

【0050】

次に、ステップ S202 では、ステップ S200 で取得した用語について、当該専門用語の文字列に含まれる部分文字列から、各部分文字列の個数を表す特徴ベクトルを生成する。

【0051】

ステップ S204 では、ステップ S202 で生成された特徴ベクトルと、判定ルール 240 とに基づいて、意味コード k 毎に、判定ルール 240 のスコア関数のパラメータ w_k の値に従って、スコア関数 $f_k(^x; ^w_k)$ を計算し、計算の結果、スコア関数 $f_k(^x_m; ^w_k)$ の値が最大になる意味コード k を、ステップ S200 で取得した用語に適した意味コードとして選択することで、ステップ S200 で取得した用語の意味コードを判定する。

【0052】

ステップ S206 では、ステップ S204 による判定で得られた用語の意味コードを出力部 250 により出力し、処理を終了する。

【0053】

以上説明したように、第 1 の実施の形態に係る用語意味判定装置によれば、用語に含まれる部分文字列から、特徴ベクトルを生成し、生成された特徴ベクトルと、予め学習された判定ルールとに基づいて、用語の意味コードを判定することで、精度よく専門用語の意味コードを判定することができる。

【0054】

< 本発明の第 2 の実施の形態に係る用語意味学習装置の構成 >

【0055】

次に、本発明の第 2 の実施の形態に係る用語意味学習装置の構成について説明する。なお、第 1 の実施の形態の用語意味学習装置 100 と同様の構成となる部分については、同一符号を付して説明を省略する。

【0056】

上記図 1 に示すように、本発明の第 2 の実施の形態に係る用語意味学習装置 100 は、CPU と、RAM と、後述する用語意味学習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。この用語意味学習装置 100 は、機能的には図 1 に示すように入力部 10 と、演算部 20 とを備えている。

【0057】

第 2 の実施の形態に係る入力部 10 は、第 1 の実施の形態と同様に、専門用語の意味コードが予め付与された複数の専門用語を、訓練データ集合として受け付け、訓練データ集合データベース 28 に格納する。

10

20

30

40

50

【 0 0 5 8 】

第 2 の実施の形態に係る演算部 2 0 は、訓練データ集合データベース 2 8 と、特徴ベクトル生成部 3 0 と、判定ルール生成部 3 2 と、判定ルール 4 0 とを含んで構成されている。

【 0 0 5 9 】

第 2 の実施の形態に係る特徴ベクトル生成部 3 0 は、入力部 1 0 で受け付けた複数の専門用語の各々について、当該専門用語に含まれる文字、及び当該専門用語の文字列に含まれる部分文字列から、各文字の個数と、各部分文字列の個数とを表す特徴ベクトルを生成する。なお、文字の有無、及び各部分文字列の有無を表す特徴ベクトルを生成してもよい。

10

【 0 0 6 0 】

第 2 の実施の形態に係る特徴ベクトル生成部 3 0 では、専門用語を構成する文字、例えば、文字ベースの $uni-gram$ と、専門用語を構成する部分文字列、例えば、 $bi-gram$ 又は $tri-gram$ などを、特徴量とし、用語に含まれる各特徴量の個数を表す特徴ベクトルを生成する。例えば、図 6 に示すような、「ピロリ菌」という専門用語に含まれる文字と、専門用語の文字列に含まれる部分文字列との特徴量のインデックスを仮定する。このとき、専門用語「ピロリ菌」は、 $i \in \{u_1, u_2, u_3, u_4, b_1, b_2, b_3, b_4, b_5, t_1, t_2, t_3, t_4\}$ のときに $x_i = 1$ 、 i が、「ピロリ菌」に含まれない文字、又は部分文字列の特徴量のときに $x_i = 0$ の要素からなる特徴ベクトル $x = (x_1, \dots, x_i, \dots, x_I)$ で表すことができる。

20

【 0 0 6 1 】

なお、第 2 の実施の形態に係る用語意味学習装置の他の構成及び作用は、第 1 の実施の形態の用語意味学習装置 1 0 0 と同様であるため詳細な説明を省略する。

【 0 0 6 2 】

< 本発明の第 2 の実施の形態に係る用語意味判定装置の構成 >

【 0 0 6 3 】

次に、本発明の第 2 の実施の形態に係る用語意味判定装置の構成について説明する。なお、第 1 の実施の形態の用語意味判定装置 2 0 0 と同様の構成となる部分については、同一符号を付して説明を省略する。

【 0 0 6 4 】

上記図 3 に示すように、本発明の第 2 の実施の形態に係る用語意味判定装置 2 0 0 は、CPU と、RAM と、後述する用語意味判定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。この用語意味判定装置 2 0 0 は、機能的には図 3 に示すように入力部 2 1 0 と、演算部 2 2 0 と、出力部 2 5 0 とを備えている。

30

【 0 0 6 5 】

第 2 の実施の形態に係る演算部 2 2 0 は、特徴ベクトル生成部 2 3 0 と、意味コード判定部 2 3 2 と、判定ルール 2 4 0 とを含んで構成されている。

【 0 0 6 6 】

第 2 の実施の形態に係る特徴ベクトル生成部 2 3 0 は、入力部 2 1 0 で受け付けた用語の文字列に含まれる文字及び部分文字列から、第 1 の実施の形態に係る特徴ベクトル生成部 3 0 と同様に、各文字の個数及び各部分文字列の個数とを表す特徴ベクトルを生成する。なお、各文字の有無及び各部分文字列の有無を表す特徴ベクトルを生成してもよい。

40

【 0 0 6 7 】

なお、第 2 の実施の形態に係る用語意味判定装置の他の構成及び作用は、第 1 の実施の形態の用語意味判定装置 2 0 0 と同様であるため詳細な説明を省略する。

【 0 0 6 8 】

以上説明したように、第 2 の実施の形態に係る用語意味学習装置によれば、複数の専門用語の各々について、用語に含まれる文字及び部分文字列から、特徴ベクトルを生成し、複数の専門用語の各々について生成された特徴ベクトルと、複数の専門用語の各々に付与

50

された意味コードとに基づいて、意味コードを判定するための判定ルールを生成することで、精度よく専門用語の意味コードを判定することができる。

【 0 0 6 9 】

また、第 2 の実施の形態に係る用語意味判定装置によれば、用語に含まれる文字及び部分文字列から、特徴ベクトルを生成し、生成された特徴ベクトルと、予め学習された判定ルールとに基づいて、用語の意味コードを判定することで、精度よく専門用語の意味コードを判定することができる。

【 0 0 7 0 】

< 本発明の第 3 の実施の形態に係る用語意味学習装置の構成 >

【 0 0 7 1 】

10

次に、本発明の第 3 の実施の形態に係る用語意味学習装置の構成について説明する。なお、第 2 の実施の形態の用語意味学習装置 1 0 0 と同様の構成となる部分については、同一符号を付して説明を省略する。

【 0 0 7 2 】

上記図 1 に示すように、本発明の第 3 の実施の形態に係る用語意味学習装置 1 0 0 は、CPU と、RAM と、後述する用語意味学習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。この用語意味学習装置 1 0 0 は、機能的には図 1 に示すように入力部 1 0 と、演算部 2 0 とを備えている。

【 0 0 7 3 】

20

第 3 の実施の形態に係る演算部 2 2 0 は、訓練データ集合データベース 2 8 と、特徴ベクトル生成部 3 0 と、判定ルール生成部 3 2 と、判定ルール 4 0 とを含んで構成されている。

【 0 0 7 4 】

第 3 の実施の形態に係る特徴ベクトル生成部 3 0 は、入力部 1 0 で受け付けた複数の専門用語の各々について、当該専門用語に含まれる文字、及び当該専門用語の文字列に含まれる部分文字列から、各文字の個数と、各部分文字列の個数と、当該専門用語に含まれる各文字の文字種とを表す特徴ベクトルを生成する。なお、各文字の有無及び各部分文字列の有無を表す特徴ベクトルを生成してもよい。

【 0 0 7 5 】

30

第 3 の実施の形態に係る特徴ベクトル生成部 3 0 では、専門用語を構成する文字、例えば、文字ベースの `uni-gram` と、専門用語を構成する部分文字列、例えば、`bi-gram` 又は `tri-gram` などの特徴量とした用語に含まれる各特徴量の個数と、専門用語を構成する各文字の文字種、例えば、ひらがな、カタカナ、又はアルファベットなどを特徴量とした用語に含まれる各特徴量の個数とを組み合わせた特徴ベクトルを生成する。例えば、図 7 に示すような、「頸部リンパ節腫脹」という専門用語に含まれる文字、及び専門用語の文字に含まれる部分文字列の特徴量と、専門用語に含まれる各文字の文字種の特徴量とを組み合わせたインデックスを仮定する。ここで、 c_t は専門用語に含まれる t 番目の文字を表し、 s_t は専門用語に含まれる t 番目の文字の文字種を表す。「頸部リンパ節腫脹」であれば、 c_t については、 $c_1 = \text{頸}$ 、 $c_2 = \text{部}$ 、 $c_3 = \text{リ}$ 、 $\dots$ と表すことができ、 s_t については、 $s_1 = s_2 = \text{"CJK unified Ideographs (漢字)"}$ 、 $s_3 = \text{"Katakana (カタカナ)"}$ 、等と表すことができる。なお、文字種の `N-gram` を特徴量として、用語に含まれる各特徴量の個数を表す特徴ベクトルを生成してもよい。

40

【 0 0 7 6 】

なお、第 3 の実施の形態に係る用語意味判定装置の他の構成及び作用は、第 2 の実施の形態の用語意味学習装置 1 0 0 と同様であるため詳細な説明を省略する。

【 0 0 7 7 】

< 本発明の第 3 の実施の形態に係る用語意味判定装置の構成 >

【 0 0 7 8 】

次に、本発明の第 3 の実施の形態に係る用語意味判定装置の構成について説明する。な

50

お、第２の実施の形態の用語意味判定装置２００と同様の構成となる部分については、同一符号を付して説明を省略する。

【００７９】

上記図３に示すように、本発明の第３の実施の形態に係る用語意味判定装置２００は、ＣＰＵと、ＲＡＭと、後述する用語意味判定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。この用語意味判定装置２００は、機能的には図３に示すように入力部２１０と、演算部２２０と、出力部２５０とを備えている。

【００８０】

第３の実施の形態に係る演算部２２０は、特徴ベクトル生成部２３０と、意味コード判定部２３２と、判定ルール２４０とを含んで構成されている。

10

【００８１】

なお、第３の実施の形態に係る用語意味判定装置の他の構成及び作用は、第２の実施の形態の用語意味判定装置２００と同様であるため詳細な説明を省略する。

【００８２】

以上説明したように、第３の実施の形態に係る用語意味学習装置によれば、複数の専門用語の各々について、用語に含まれる文字及び部分文字列、並びに文字の文字種から、特徴ベクトルを生成し、複数の専門用語の各々について生成された特徴ベクトルと、複数の専門用語の各々に付与された意味コードとに基づいて、意味コードを判定するための判定ルールを生成することで、精度よく専門用語の意味コードを判定することができる。

20

【００８３】

また、第３の実施の形態に係る用語意味判定装置によれば、用語に含まれる文字及び部分文字列から、特徴ベクトルを生成し、生成された特徴ベクトルと、予め学習された判定ルールとに基づいて、用語の意味コードを判定することで、精度よく専門用語の意味コードを判定することができる。

【００８４】

< 本発明の第４の実施の形態に係る用語意味学習装置の構成 >

【００８５】

次に、本発明の第４の実施の形態に係る用語意味学習装置の構成について説明する。なお、第２の実施の形態の用語意味学習装置１００と同様の構成となる部分については、同一符号を付して説明を省略する。

30

【００８６】

図８に示すように、本発明の第４の実施の形態に係る用語意味学習装置３００は、ＣＰＵと、ＲＡＭと、後述する用語意味学習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。この用語意味学習装置１００は、機能的には図８に示すように入力部３１０と、演算部３２０とを備えている。

【００８７】

入力部３１０は、専門用語の意味コードが予め付与された複数の専門用語と、専門用語の同義語及び類義語を格納した辞書とを、受け付ける。

40

【００８８】

演算部３２０は、同義語類義語意味コード付与部３２６と、訓練データ集合データベース２８と、特徴ベクトル生成部３０と、判定ルール生成部３２と、判定ルール４０とを含んで構成されている。

【００８９】

同義語類義語意味コード付与部３２６は、入力部３１０で受け付けた、専門用語の意味コードが予め付与された複数の専門用語と辞書とに基づいて、意味コードが付与された専門用語の各々について、当該専門用語の同義語又は類義語となる専門用語の各々に、同じ意味コードを付与する。そして、入力部３１０で受け付けた、意味コードが予め付与された専門用語、及び意味コードが付与された同義語または類義語となる専門用語を、訓練デ

50

ータ集合として、訓練データ集合データベース 28 に格納する。

【0090】

なお、第3の実施の形態に係る用語意味学習装置の他の構成及び作用は、第2の実施の形態の用語意味学習装置100と同様であるため詳細な説明を省略する。

【0091】

<本発明の第4の実施の形態に係る用語意味判定装置の構成>

【0092】

次に、本発明の第4の実施の形態に係る用語意味判定装置の構成について説明する。なお、第2の実施の形態の用語意味判定装置200と同様の構成となる部分については、同一符号を付して説明を省略する。

10

【0093】

上記図3に示すように、本発明の第4の実施の形態に係る用語意味判定装置200は、CPUと、RAMと、後述する用語意味判定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。この用語意味判定装置200は、機能的には図3に示すように入力部210と、演算部220と、出力部250とを備えている。

【0094】

第4の実施の形態に係る演算部220は、特徴ベクトル生成部230と、意味コード判定部232と、判定ルール240とを含んで構成されている。

【0095】

20

なお、第4の実施の形態に係る用語意味判定装置の他の構成及び作用は、第2の実施の形態の用語意味判定装置200と同様であるため詳細な説明を省略する。

【0096】

<本発明の第4の実施の形態に係る用語意味学習装置の作用>

【0097】

次に、本発明の第4の実施の形態に係る用語意味学習装置300の作用について説明する。入力部10において、専門用語の意味コードが予め付与された複数の専門用語、及び同義語及び類似語を格納した辞書を受け付けると、第4の実施の形態に係る用語意味学習装置300は、図9に示す用語意味学習処理ルーチンを実行する。なお、第4の実施の形態に係る用語意味学習装置の作用について、第2の実施の形態と同様の作用となる部分については、同一符号を付して説明を省略する。

30

【0098】

ステップS300では、同義語及び類似語を格納した辞書に基づいて、ステップS100で取得した複数の専門用語の各々について、当該専門用語の同義語又は類義語となる専門用語に対して、当該専門用語と同じ意味コードを付与する。

【0099】

なお、第4の実施の形態に係る用語意味学習装置の他の作用については、第2の実施の形態と同様であるため、詳細な説明を省略する。

【0100】

以上説明したように、第4の実施の形態に係る用語意味学習装置によれば、辞書に格納されている同義語及び類義語に対して、意味コードを付与することで、意味コードを付与した訓練データを増やすことができる。

40

【0101】

また、第4の実施の形態に係る用語意味判定装置によれば、用語に含まれる文字及び部分文字列から、特徴ベクトルを生成し、生成された特徴ベクトルと、予め学習された判定ルールとに基づいて、用語の意味コードを判定することで、精度よく専門用語の意味コードを判定することができる。

【0102】

<実験結果>

【0103】

50

上記第3の実施の形態に係る手法を適用して、国立情報学研究所（NII）主催の評価型ワークショップNTCIR-11のMedNLP2タスク（<http://mednlp.jp/ntcir11/>）で配布された疑似医療文書のベンチマークセットを用いて実験を行った。このベンチマークセットには、訓練データとテストデータが含まれており、それぞれのデータには、疑似医療文書中に、意味コードとしてICDコードが付与された病名及び症状を表す専門用語を含むものである。ICDコードは、病態管理および情報共有のためのコードであり、医療用語の標準化のために開発されたコードである。ICDコードを利用することで、様々な省略形や漢字、かな表記の違いなど、同一の意味をもつ異表現の用語を把握、管理することができる。ICD10対応標準病名マスター（<http://www2.medis.or.jp/stdcd/byomei/>）には数千種類のICDコードが定義されている。

10

【0104】

性能評価では、ICDコードを意味コードとし、疑似医療文書に含まれる病名及び症状を表す専門用語にICDコードを付与する問題に対して、上記第3の実施の形態に係る手法を適用した場合に得られるICDコードの判定精度を調べた。

【0105】

本実験では、上記ベンチマークセットの訓練データと、ICD10対応標準病名マスターに含まれる病名基本テーブルに含まれる専門用語とを訓練データ集合データベース28に格納した。また、ICD10対応標準病名マスターの検索テーブルに含まれる専門用語も訓練データ集合データベース28に格納した。また、ICD10対応標準病名マスターの病名基本テーブルに含まれる専門用語と同義又は類義の専門用語について、当該病名基本テーブルに含まれる専門用語と同一のICDコードを付与して訓練データ集合データベース28に格納した。さらに、ライフサイエンス辞書（<http://lsd.pharm.kyoto-u.ac.jp/ja/index.html>）、T辞書（<http://www.tdic.co.jp/>）、及びUMLSメタソーラス（<http://www.nlm.nih.gov/research/umls/>）から取り出したMedDRA辞書の各々から、ICD10対応標準病名マスターに含まれる専門用語と同義又は類義の専門用語を抽出し、当該ICD10対応標準病名マスターに含まれる専門用語と同一のICDコードを付与して訓練データ集合データベース28に格納した。

20

【0106】

本実験では、意味コード判定部232で用いる判定ルールを生成するために、判定ルール生成部32において、ロジスティック回帰モデルを適用して、スコア関数 $f_k(\hat{x}; \hat{w}_k)$ のパラメータ値を推定した。

30

【0107】

表1に、上記第3の実施の形態に係る手法で判定ルールを生成して、ベンチマークセットのテストデータに含まれる2134個の病名及び症状を表す専門用語に適用した場合の意味コードの正解率と、正しく意味コードが推定された専門用語の個数（正解数）を示す。表1において、ADOC、MEDIS、及びEXDICは訓練データ集合データベース28に格納された用語の種類を表している。ADICはベンチマークセットの訓練データに格納されていた用語を、MEDISはICD10対応標準病名マスターの基本病名テーブルと検索テーブルから抽出した用語を、EXDICはライフサイエンス辞書、T辞書、及びMedDRA辞書から抽出した用語を表す。表1から、より多くの専門用語を辞書から集めて判定ルールの生成に利用することで意味コードの推定精度が向上することがわかる。

40

【0108】

【表 1】

訓練データ DB	正解率	正解数
ADOC	0.657	1403
MEDIS	0.551	1176
MEDIS + EXDIC	0.578	1235
ADOC + MEDIS	0.789	1685
ADOC + MEDIS + EXDIC	0.792	1691

10

【0109】

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。

【0110】

例えば、上述した第1～第4の実施の形態では、特定の専門分野において唯一の意味が定義されている専門用語から、意味コードを判定するため判定ルールを生成する用語意味学習装置を例に説明したが、これに限定されるものではなく、意味コードが予め付与された用語を用いるのであれば、どのような分野の用語であっても判定ルールを生成し、意味コードを判定することができる。例えば、用語意味学習装置で、ある一定の分野（自然科学、経済等の分野）の用語について、判定ルールを生成し、用語意味判定装置で、生成された判定ルールを用いて、当該分野の用語について、意味コードを判定することができる。

20

【0111】

また、用語意味学習装置と用語意味判定装置とを別々の装置で構成する場合を例に説明したが、これに限定されるものではなく、用語意味学習装置と用語意味判定装置とを一つの装置で構成するようにしてもよい。

【符号の説明】

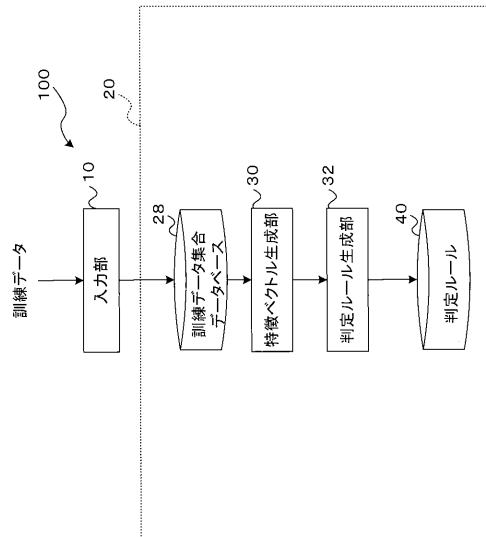
【0112】

- 10、210、310 入力部
- 20、220、320 演算部
- 28 訓練データ集合データベース
- 30、230 特徴ベクトル生成部
- 32 判定ルール生成部
- 40、240 判定ルール
- 100、300 用語意味学習装置
- 200 用語意味判定装置
- 232 意味コード判定部
- 250 出力部
- 326 同義語類義語意味コード付与部

30

40

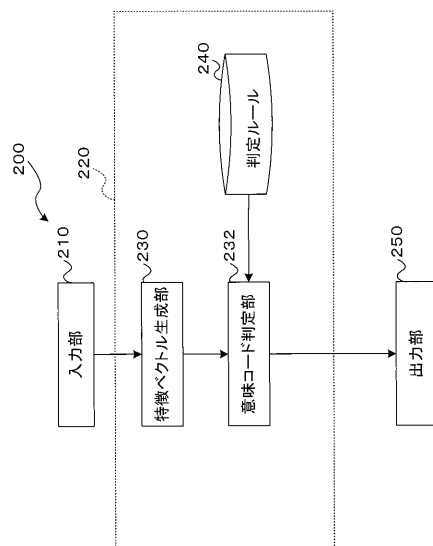
【図 1】



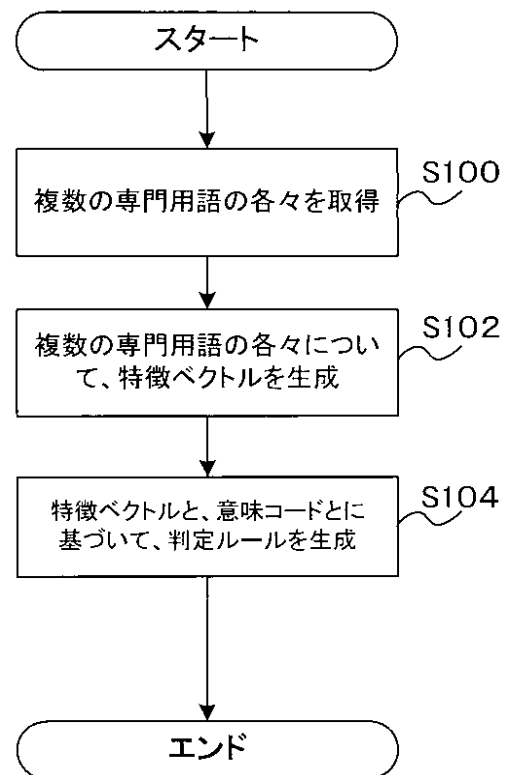
【図 2】

- 「_ピ」: b_1
- 「ピロ」: b_2
- 「ロリ」: b_3
- 「リ菌」: b_4
- 「菌_」: b_5
- 「_ピロ」: t_1
- 「ピロリ」: t_2
- 「ロリ菌」: t_3
- 「リ菌_」: t_4

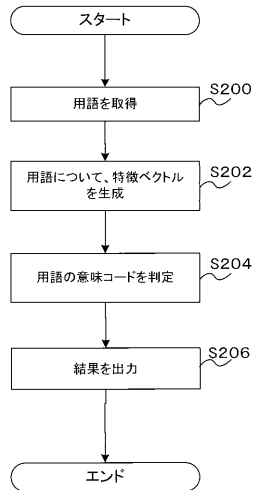
【図 3】



【図 4】



【図 5】



【図 6】

- 「ピ」: u_1
- 「口」: u_2
- 「リ」: u_3
- 「菌」: u_4
- 「_ピ」: b_1
- 「ピ口」: b_2
- 「口リ」: b_3
- 「リ菌」: b_4
- 「菌_」: b_5
- 「_ピ口」: t_1
- 「ピ口リ」: t_2
- 「口リ菌」: t_3
- 「リ菌_」: t_4

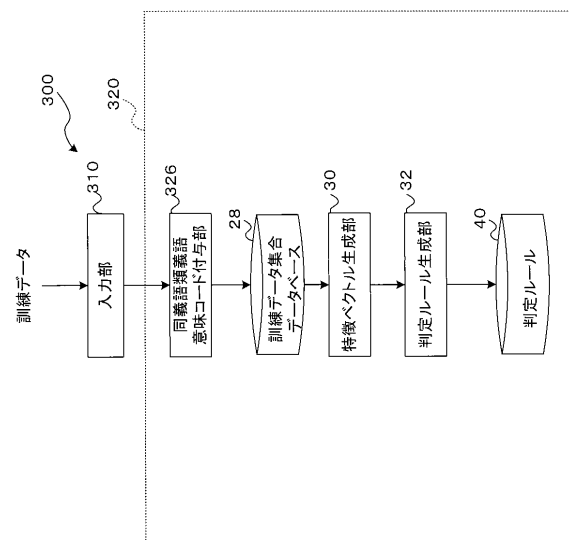
【図 7】

Ex. 頸部リンパ節腫脹

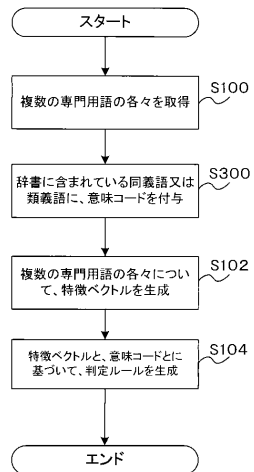
$$x = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 3 \\ \vdots \end{bmatrix}$$

$\leftarrow c_t = \text{頸}$
 $\leftarrow c_t = \text{部}$
 $\leftarrow c_{t-1}c_t = \text{頸部}$
 $\leftarrow c_{t-1}c_t = \text{部リ}$
 $\leftarrow s_t = \text{"katakana"}$

【図 8】



【図 9】



フロントページの続き

審査官 長 由紀子

- (56)参考文献 特開2008-021139(JP,A)
特開2009-230270(JP,A)
米国特許出願公開第2013/0151443(US,A1)
山田 寛康 外2名,単語の部分文字列を考慮した専門用語抽出と分類,情報処理学会研究報告
 ,日本,社団法人情報処理学会,2000年11月22日,第2000巻第107号,p.77-84
- (58)調査した分野(Int.Cl.,DB名)
G06F 17/20-30