(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
2023年12月7日(07.12.2023)

WIPO | PCT

(10) 国際公開番号
WO 2023/233633 A1

(51) 国際特許分類:
G06F 40/166 (2020.01) *G06F 40/253* (2020.01)
G06F 40/216 (2020.01) *G06F 40/44* (2020.01)
G06F 40/232 (2020.01)

(21) 国際出願番号： PCT/JP2022/022525

(22) 国際出願日： 2022年6月2日(02.06.2022)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 富士通株式会社 (**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

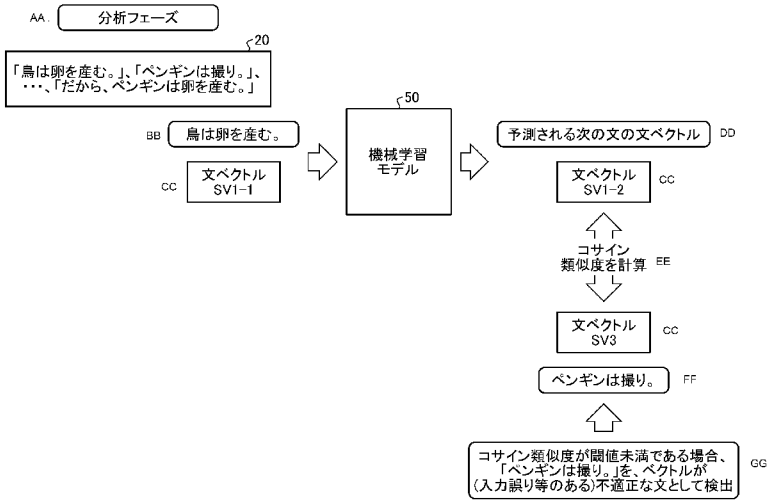
(72) 発明者: 片岡 正弘 (**KATAOKA, Masahiro**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP). 松村 量 (**MATSUMURA, Ryo**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP). 尾上 聡 (**ONOUE, Satoshi**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

(74) 代理人: 弁理士法人酒井国際特許事務所 (**SAKAI INTERNATIONAL PATENT OFFICE**); 〒1000013 東京都千代田区霞が関３丁目８番１号 虎の門三井ビルディング 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,

(54) **Title:** INFORMATION PROCESSING PROGRAM, INFORMATION PROCESSING METHOD, AND INFORMATION PROCESSING DEVICE

(54) 発明の名称： 情報処理プログラム、情報処理方法および情報処理装置



20 "Bird lays egg.", "Penguin takes photo.", "Therefore, penguin lays egg."
50 Machine-learning model
AA Analysis phase
BB Bird lays egg
CC Sentence vectors
DD Sentence vectors of predicted next sentence
EE Compute cosine similarity
FF Penguin takes photo.
GG When cosine similarity is smaller than threshold, detect "Penguin takes photo."as sentence of improper vectors (having input error, etc.)

(57) **Abstract:** This information processing device calculates vectors of a plurality of sentences having a relationship in front and back sentences in a plurality of continuous sentences. The information processing device generates a machine-learning model which predicts sentence vectors of a sentence that is input after a certain sentence when the sentence vectors of the certain sentence are input to the machine-learning model by inputting, in order, the plurality of sentence vectors to the machine-learning model and training the machine-learning model. The information processing device cal-

WO 2023/233633 A1

[続葉有]

CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, 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, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(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, SC, SD, SL, ST, 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：

一　国際調査報告（条約第21条(3)）

culates vectors of a first sentence and vectors of a second sentence that follows the first sentence. The information processing device calculates vectors of a sentence that is predicted to follow the first sentence by inputting the vectors of the first sentence to the machine-learning model, and determines whether the vectors of the second sentence are proper.

(57) 要約：情報処理装置は、連続する複数の文であって、前後の文に関係性を有する複数の文のベクトルをそれぞれ算出する。情報処理装置は、複数の文のベクトルを順番に機械学習モデルに入力して訓練することで、機械学習モデルにある文のベクトルを入力した際に、ある文の次に入力される文の文ベクトルを予測する機械学習モデルを生成する。情報処理装置は、第１文のベクトルと、第１文に続く第２文のベクトルとを算出する。情報処理装置は、第１文のベクトルを機械学習モデルに入力することで第１文に続くと予測される文のベクトルを算出し、第２文のベクトルが適正であるか否かを判定する。

明　細　書

発明の名称 ：

情報処理プログラム、情報処理方法および情報処理装置

技術分野

[0001]　本発明は、情報処理プログラム等に関する。

背景技術

[0002]　近年、文のベクトルを算出し、算出したベクトルを利用して、他言語への翻訳、データベースの検索等の各種処理を実行するサービスが提供されている。しかし、ユーザに指定される文自体に入力誤り等が存在すると、文のベクトルを精度よく算出することができず、翻訳、検索等の処理に誤りが生じる場合がある。

[0003]　たとえば、適正な文「その機能が特徴である」と、入力誤りの文「その昨日が特徴である」とは、相互に大きく意味の異なる文となり、各文のベクトルも大きく異なる。

[0004]　文の入力誤りを修正する従来技術として、修正履歴から、入力誤りと、その修正文とのペアのデータセットを用いて、学習モデルを訓練しておき、訓練した学習モデルに対象となる文を入力することで、対象となる文の入力誤りを修正する従来技術がある。

先行技術文献

特許文献

[0005] 特許文献1：特開２０１９－１０１９９３号公報

非特許文献

[0006] 非特許文献1：三木一弘、他“BERTを用いた英文空所補充問題の一解法”岡山大学工学部情報系学科、DEIM2020　G2-4　(day1　p47)

非特許文献2：田中佑、他“Wikipediaの修正履歴を用いた日本語入力誤りデータセットの構築”京都大学　大学院情報学科研究科、言語処理学会、第26回年次大会、発表論文集、2020年3月

発明の概要

発明が解決しようとする課題

[0007] しかしながら、上述した従来技術では、文章の一部の単語などがマスクされた空所補充する技術であり、複数の単語で構成される文の空所を補充する単語の精度は高いものの、複数の文で構成される文章の空所を補充する文に関する高精度化の記述は少なく、かつ、入力誤りを含む文を検出するものではなかった。また、従来技術では、誤字、脱字等の入力誤りを修正できるものの、誤変換の入力誤りを正しく修正できないケースが多かった。

[0008] １つの側面では、本発明は、複数の文で構成される文章の空所を補充する文の推定や、入力誤りを含む文を検出することができる情報処理プログラム、情報処理方法および情報処理装置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] 第１の案では、コンピュータに次の処理を実行させる。コンピュータは、連続する複数の文であって、前後の文に関係性を有する複数の文のベクトルをそれぞれ算出する。コンピュータは、複数の文のベクトルを順番に機械学習モデルに入力して訓練することで、機械学習モデルにある文のベクトルを入力した際に、ある文の次に入力される文の文ベクトルを予測する機械学習モデルを生成する。コンピュータは、第１文のベクトルと、第１文に続く第２文のベクトルとを算出する。コンピュータは、第１文のベクトルを前記機械学習モデルに入力することで第１文に続くと予測される文のベクトルを算出し、第２文のベクトルが適正であるか否かを判定する。

発明の効果

[0010] 複数の文で構成される文章の空所を補充する文の推定や、入力誤りを含む文を検出することができる。

図面の簡単な説明

[0011] [図1]図１は、本実施例に係る情報処理装置の学習フェーズの処理を説明するための図である。

[図2]図２は、本実施例に係る情報処理装置の分析フェーズの処理を説明するための図である。

[図3]図３は、本実施例に係る情報処理装置の構成を示す機能ブロック図である。

[図4]図４は、単語ベクトル辞書のデータ構造の一例を示す図である。

[図5A]図５Ａは、文ベクトルを算出する処理を説明するための図（１）である。

[図5B]図５Ｂは、文ベクトルを算出する処理を説明するための図（２）である。

[図6]図６は、文転置インデックスを生成する処理を説明するための図である。

[図7]図７は、本実施例に係る情報処理装置の学習フェーズの処理手順を示すフローチャートである。

[図8]図８は、本実施例に係る情報処理装置の分析フェーズの処理手順を示すフローチャートである。

[図9]図９は、情報処理装置のその他の処理を説明するための図（１）である。

[図10]図１０は、情報処理装置のその他の処理を説明するための図（２）である。

[図11]図１１は、実施例の情報処理装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。

発明を実施するための形態

[0012] 以下に、本願の開示する情報処理プログラム、情報処理方法および情報処理装置の実施例を図面に基づいて詳細に説明する。なお、この実施例によりこの発明が限定されるものではない。

実施例

[0013] 本実施例に係る情報処理装置の処理について説明する。情報処理装置は、学習フェーズの処理を実行した後に、分析フェーズの処理を実行する。図１

は、本実施例に係る情報処理装置の学習フェーズの処理を説明するための図である。

[0014] 学習フェーズにおいて、情報処理装置は、教師データ１４１に含まれる複数の文章を用いて、機械学習モデル５０の学習を実行する（機械学習モデル５０を訓練する）。機械学習モデル５０は、BERT（Pre-training of Deep Bidirectional Transformers for Language Understanding）、Next Sentence Prediction、Transformers等のＮＮ（Neural Network）である。

[0015] 教師データ１４１に含まれる文章には、複数の文が含まれる。複数の文は、前後の文に所定の関係性を有する。各文は、帰納法や演繹法の三段論法等に基づいて予め設定された文である。

[0016] たとえば、文章１０ａには、先頭から順に、文「鳥は卵を産む。」、文「ペンギンは鳥。」、・・・、文「だから、ペンギンは卵を産む。」が含まれる。文章１０ｂには、先頭から順に、文「鳥は卵から生まれる。」、文「鳩は鳥の仲間である。」、・・・、文「従って、鳩は卵から生まれる。」が含まれる。

[0017] 情報処理装置は、文章１０ａ，１０ｂ、その他の文章に含まれる各文の文ベクトルを算出する。たとえば、情報処理装置は、文に対して形態素解析を実行して単語に分割し、各単語のベクトルを積算することで、文ベクトルを算出する。

[0018] 文章１０ａの文「鳥は卵を産む。」の文ベクトルを「ＳＶ１－１」とする。文「ペンギンは鳥」の文ベクトルを「ＳＶ１－２」とする。文「だから、ペンギンは卵を産む。」の文ベクトルを「ＳＶ１－３」とする。

[0019] 文章１０ｂの文「鳥は卵から生まれる。」の文ベクトルを「ＳＶ２－１」とする。文「鳩は鳥の仲間である。」の文ベクトルを「ＳＶ２－２」とする。文「従って、鳩は卵から生まれる。」の文ベクトルを「ＳＶ２－３」とする。

[0020] 情報処理装置は、文章に含まれる先頭の文のベクトルから順番に、機械学習モデル５０に入力する処理を繰り返し実行する。たとえば、情報処理装置

は、文ベクトル「ＳＶ１－１」、「ＳＶ１－２」、・・・、「ＳＶ１－３」の順に、機械学習モデル５０に文ベクトルを入力する。情報処理装置は、文ベクトル「ＳＶ２－１」、「ＳＶ２－２」、・・・、「ＳＶ２－３」の順に、機械学習モデル５０に文ベクトルを入力する。

[0021] 情報処理装置が、上記の学習フェーズの処理を実行することで、ある第１文の文ベクトルが入力された場合に、第１文の次の第２文の文ベクトルを予測する機械学習モデル５０が生成される。

[0022] 図２は、本実施例に係る情報処理装置の分析フェーズの処理を説明するための図である。分析フェーズにおいて、情報処理装置は、訓練済みの機械学習モデル５０を用いて、処理対象の文章に含まれる文ベクトルを算出し、コサイン類似度などにより、不適正な文を検出する。

[0023] 図２の説明では、入力誤り等を含む処理対象の文章を文章２０とする。文章２０には、先頭から順に、文「鳥は卵を産む。」、文「ペンギンは撮り。」、・・・、文「だから、ペンギンは卵を産む。」により構成される。文「ペンギンは撮り。」は、教師データ１４１の文章１０aに含まれる正しい文「ペンギンは鳥。」に対して、単語「鳥」の同音異義語「撮り」の入力誤りを含む文である。

[0024] 情報処理装置は、文「鳥は卵を産む。」の文ベクトル「ＳＶ１－１」を算出し、算出した文ベクトル「ＳＶ１－１」を、機械学習モデル５０に入力することで、文「鳥は卵を産む。」の次の文の文ベクトルを予測する。図２に示す例では、機械学習モデル５０によって、文「鳥は卵を産む。」の次の文の文ベクトルとして、「ＳＶ１－２」が予測されている。

[0025] 情報処理装置は、文章２０に含まれる文であって、文「鳥は卵を産む。」の次の文「ペンギンは撮り」の文ベクトル「ＳＶ３」を算出する。

[0026] 情報処理装置は、機械学習モデル５０によって予測された次の文の文ベクトル「ＳＶ１－２」と、文章２０に含まれる文であって、文「鳥は卵を産む。」の次の文「ペンギンは撮り」の文ベクトル「ＳＶ３」とのコサイン類似度を算出する。

[0027] 情報処理装置は、文章１０ａに含まれる文であって、文「鳥は卵を産む。」の次の文「ペンギンは鳥。」は、コサイン類似度が閾値未満の場合として、正しい（以下、適正な、と表記する）文であると判定する。一方、情報処理装置は、文章２０に含まれる文であって、文「鳥は卵を産む。」の次の文「ペンギンは撮り。」は、コサイン類似度が閾値未満の場合として、入力誤り等を含む不適正な文であると判定する。

[0028] 上記のように、情報処理装置は、教師データ１４１に含まれる文章の各文のベクトルを順番に機械学習モデル５０に入力することで、ある第１文の文ベクトルが入力された場合に、第１文の次の第２文の文ベクトルを予測する機械学習モデル５０を生成する。情報処理装置は、生成した機械学習モデルに、処理対象の文章の文の文ベクトルを入力し、次の文の文ベクトルを予測し、予測した文ベクトルを基にして、処理対象の文章から、入力誤りのある文を検出する。すなわち、処理対象の文章に含まれる各文から、入力誤り等を含み、不適正な文ベクトルを持つ文を検出することができる。

[0029] なお、情報処理装置は、図２の処理において、文「ペンギンは撮り」が不適正な文ベクトルの文であると判定した場合に、機械学習モデル５０によって予測された文ベクトルＳＶ１－２を基にして、適正な文ベクトルの文「ペンギンは鳥。」をＤＢ（Data Base）などから検索して、正しい修正候補として表示装置に出力（以下、適正化と表記する）してもよい。

[0030] 更に、情報処理装置は、単語単位のベクトルの順番を学習した他の機械学習モデルに、不適正な文ベクトルを検出した文「ペンギンは撮り」を構成する複数の単語「ペンギン」、「は」、「撮り。」の各単語ベクトルを算出し、乖離した単語「撮り。」の入力誤り等を適正化してもよい。

[0031] 次に、図１及び図２で説明した処理を実行する情報処理装置の構成例について説明する。図３は、本実施例に係る情報処理装置の構成を示す機能ブロック図である。図３に示すように、情報処理装置１００は、通信部１１０、入力部１２０、表示部１３０、記憶部１４０、制御部１５０を有する。

[0032] 通信部１１０は、有線又は無線で外部装置等に接続され、外部装置等との

間で情報の送受信を行う。たとえば、通信部１１０は、ＮＩＣ（Network Interface Card）等によって実現される。通信部１１０は、図示しないネットワークに接続されていてもよい。

[0033] 入力部１２０は、各種の情報を、情報処理装置１００に入力する入力装置である。入力部１２０は、キーボードやマウス、タッチパネル等に対応する。たとえば、ユーザは、入力部１２０を操作して、文章のデータ等を入力してもよい。

[0034] 表示部１３０は、制御部１５０から出力される情報を表示する表示装置である。表示部１３０は、液晶ディスプレイ、有機ＥＬ（Electro Luminescence）ディスプレイ、タッチパネル等に対応する。たとえば、入力誤りのある文が、表示部１３０に表示される。

[0035] 記憶部１４０は、機械学習モデル５０、教師データ１４１、単語ベクトル辞書１４２を有する。記憶部１４０は、たとえば、ＲＡＭ（Random Access Memory）、フラッシュメモリ（Flash Memory）等の半導体メモリ素子、または、ハードディスク、光ディスク等の記憶装置によって実現される。

[0036] 機械学習モデル５０は、図１で説明したBERT、Next Sentence Prediction、Transformers等のＮＮ等である。

[0037] 教師データ１４１は、図１で説明した教師データ１４１である。教師データ１４１に含まれる文章には、複数の文が含まれる。複数の文は、前後の文に所定の関係性を有する。各文は、帰納法や演繹法の三段論法等に基づいて予め設定された文である。

[0038] 単語ベクトル辞書１４２は、単語に割り当てられた符号、単語ベクトルを定義するテーブルである。図４は、単語ベクトル辞書のデータ構造の一例を示す図である。図４に示すように、この単語ベクトル辞書１４２は、符号、単語、単語ベクトル（１）～（７）を有する。符号は、単語に割り当てられる符号（Code）である。単語は、文字列に含まれる単語である。単語ベクトル（１）～（７）は、単語に割り当てられたベクトルである。単語ベクトルの第ｎ成分を単語ベクトル（ｎ）と表記する（ｎ＝１～７）。

[0039] ＤＢ１４３は、様々な文章を有する。文章には複数の文が含まれ、各文には複数の単語が含まれる。ＤＢ１４３は、教師データ１４１に含まれる文章を有していてもよい。

[0040] 文転置インデックス１４４は、文ベクトルと、位置ポインタとを対応付ける。位置ポインタは、文ベクトルに対応する文が存在するＤＢ１４３の位置を示す。

[0041] 図３の説明に戻る。制御部１５０は、前処理部１５１と、学習部１５２と、分析部１５３とを有する。制御部１５０は、たとえば、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing Unit）により実現される。また、制御部１５０は、例えばＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）等の集積回路により実行されてもよい。

[0042] 前処理部１５１は、各種の前処理を実行する。たとえば、前処理部１５１は、ＤＢ１４３から未処理に文を取得し、文の文ベクトルを算出する。前処理部１５１は、算出した文ベクトルと、文ベクトルに対応する文の位置ポインタとの関係を、文転置インデックス１４４に設定する。

[0043] 前処理部１５１が、文の文ベクトルを算出する処理の一例について説明する。図５Ａ及び図５Ｂは、文ベクトルを算出する処理を説明するための図である。ここでは、文「馬は人参が好きです。」の文ベクトルを算出する場合について説明する。前処理部１５１は、文「馬は人参が好きです。」に対して、形態素解析を実行することで、複数の単語に分解する。分解した各単語には、「△（スペース）」を付与する。たとえば、文１「馬は人参が好きです。」は、「馬△」、「は△」、「人参△」、「が△」、「好き△」、「です△」、「。△」に分割する。

[0044] 前処理部１５１は、分割した各単語と、単語ベクトル辞書４５とを比較することで、各単語に対応する符号を特定し、単語と置き換える。たとえば、各単語「馬△」、「は△」、「人参△」、「が△」、「好き△」、「です△」、「。△」は、それぞれ、「Ｃ１」、「Ｃ２」、「Ｃ３」、「Ｃ４」、「

Ｃ５」、「Ｃ６」、「Ｃ７」に置き換えられる。

[0045] 図５Ｂの説明に移行する。前処理部１５１は、単語ベクトル辞書４５と、各符号とを基にして、符号に割り当てられた単語ベクトル（１）～（７）を特定する。たとえば、符号「Ｃ１」の単語ベクトル（１）～（７）は、ｗｖ１－１～１－７とする。符号「Ｃ２」の単語ベクトル（１）～（７）は、ｗｖ２－１～２－７とする。符号「Ｃ３」の単語ベクトル（１）～（７）は、ｗｖ３－１～３－７とする。

[0046] 符号「Ｃ４」の単語ベクトル（１）～（７）は、ｗｖ４－１～４－７とする。符号「Ｃ５」の単語ベクトル（１）～（７）は、ｗｖ５－１～５－７とする。符号「Ｃ６」の単語ベクトル（１）～（７）は、ｗｖ６－１～６－７とする。符号「Ｃ７」の単語ベクトル（１）～（７）は、ｗｖ７－１～７－７とする。

[0047] 前処理部１５１は、要素毎に単語ベクトルを積算することで、文の文ベクトルＳＶ１を算出する。たとえば、前処理部１５１は、各単語ベクトル（１）となるｗｖ１－１～７－１を積算することで、文ベクトルＳＶ１の第１成分「ＳＶ１－１」を算出する。前処理部１５１は、各単語ベクトル（２）となるｗｖ１－２～７－２を積算することで、文ベクトルＳＶ１の第２成分「ＳＶ１－２」を算出する。各単語ベクトル（３）となるｗｖ１－３～７－３を積算することで、文ベクトルＳＶ１の第３成分「ＳＶ１－３」を算出する。

[0048] 前処理部１５１は、各単語ベクトル（４）となるｗｖ１－４～７－４を積算することで、文ベクトルＳＶ１の第４成分「ＳＶ１－４」を算出する。前処理部１５１は、各単語ベクトル（５）となるｗｖ１－５～７－５を積算することで、文ベクトルＳＶ１の第５成分「ＳＶ１－５」を算出する。前処理部１５１は、各単語ベクトル（６）となるｗｖ１－６～７－６を積算することで、文ベクトルＳＶ１の第６成分「ＳＶ１－６」を算出する。前処理部１５１は、各単語ベクトル（７）となるｗｖ１－７～７－７を積算することで、文ベクトルＳＶ１の第７成分「ＳＶ１－７」を算出する。

[0049] 前処理部１５１は、ＤＢ１４３に含まれる他の文章の各文についても、上記処理を繰り返し実行することで、各文の文ベクトルを算出する。

[0050] 前処理部１５１は、算出した各文の文ベクトルと、ＤＢ１４３の位置ポインタとを対応付けることで、文転置インデックス１４４を生成する。なお、前処理部１５１は、図６に示すようなデータ構造の文転置インデックス１４４を生成してもよい。図６は、文転置インデックスを生成する処理を説明するための図である。図６に示すように、前処理部１５１は、文ベクトルと、複数のレコードポインタと、複数の位置ポインタとを対応付け、各レコードポインタ、位置ポインタを、ＤＢ１４３の各文に対応付けてもよい。

[0051] 図３の説明に戻る。学習部１５２は、図１で説明した学習フェーズの処理を実行することで、ある第１文の文ベクトルが入力された場合に、第１文の次の第２文の文ベクトルを予測する機械学習モデル５０を生成する。

[0052] たとえば、学習部１５２は、教師データ１４１の文章に含まれる各文の文ベクトルを算出し、算出した文ベクトルを順番に機械学習モデル５０に入力することで、機械学習モデル５０の学習を実行する。学習部１５２のその他の処理は、図１で説明した処理と同様である。学習部１５２が、文の文ベクトルを算出する処理は、前処理部１５１が文の文ベクトルを算出する処理と同様である。

[0053] 分析部１５３は、図２で説明した分析フェーズの処理を実行することで、処理対象の文章に含まれる文から、文ベクトルが不適正な文を検出する。

[0054] たとえば、分析部１５３は、処理対象の文章２０を受け付けた場合に、文章２０に含まれる文の文ベクトルを算出する。分析部１５３は、文章２０に含まれる句点「。」を基にして、文章２０に含まれる文を特定する。分析部１５３が、文の文ベクトルを算出する処理は、前処理部１５１が文の文ベクトルを算出する処理と同様である。文章２０の先頭からｎ番目の文の文ベクトル「ＳＶｎ」と表記する（ｎ＝０～Ｍ）。

[0055] 分析部１５３は、文ベクトルＳＶｎを、訓練済みの機械学習モデル５０に入力して、文章２０の先頭からｎ＋１番目の文の文ベクトルＳＶｎ＋１’を

予測する。分析部１５３は、機械学習モデル５０を用いて予測した文ベクトルＳＶｎ＋１’と、文のベクトルＳＶｎ＋１とのコサイン類似度を算出する。

[0056] 分析部１５３は、文ベクトルＳＶｎ＋１’と、文ベクトルＳＶｎ＋１とのコサイン類似度が閾値以上の場合には、先頭からｎ＋１番目の文が適正な文であると判定する。一方、分析部１５３は、文ベクトルＳＶｎ＋１’と、文ベクトルＳＶｎ＋１とのコサイン類似度が閾値以上の場合には、先頭からｎ＋１番目の文が、文ベクトルの不適正な文であると判定する。

[0057] 分析部１５３は、文ベクトルの不適正な文であると判定した場合に、文ベクトルＳＶｎ＋１’と、文転置インデックス１４４とを比較して、文ベクトルＳＶｎ＋１’に対応する文の位置ポインタを特定する。分析部１５３は、位置ポインタを基にして、文ベクトルＳＶｎ＋１’に対応する文を、ＤＢ１４３から検索する。分析部１５３は、文ベクトルの不適正な文と、検索した文とを対応付けて、表示部１３０に表示させる。

[0058] 分析部１５３は、文ベクトルの不適正な文と、検索した文とを単語単位に比較して、文ベクトルの不適正な文から、入力誤りの単語を検出し、検出した単語を表示させてもよい。

[0059] 次に、本実施例に係る情報処理装置１００の処理手順の一例について説明する。図７は、本実施例に係る情報処理装置の学習フェーズの処理手順を示すフローチャートである。図７に示すように、情報処理装置１００の学習部１５２は、教師データ１４１から、未選択の文章を選択する（ステップＳ１０１）。

[0060] 学習部１５２は、選択した文章に含まれる各文の文ベクトルをそれぞれ算出し、文ベクトルとＤＢのレコードと文の位置とを対応付けた文転置インデックスを生成する（ステップＳ１０２）。学習部１５２は、選択した文章に含まれる先頭の文の文ベクトルから順に、機械学習モデル５０に入力することで、学習を実行する（ステップＳ１０３）。

[0061] 学習部１５２は、学習を継続する場合には（ステップＳ１０４，Ｙｅｓ）

、ステップＳ１０１に移行する。一方、学習部１５２は、学習を継続しない場合には（ステップＳ１０４，Ｎｏ）、学習フェーズの処理を終了する。

[0062] 図８は、本実施例に係る情報処理装置の分析フェーズの処理手順を示すフローチャートである。図８に示すように、情報処理装置１００の分析部１５３は、処理対象の文章の入力を受け付ける（ステップＳ２０１）。

[0063] 分析部１５３は、入力された文章に含まれる各文の文ベクトルをそれぞれ算出する（ステップＳ２０２）。分析部１５３は、ｎを初期値に設定する（ステップＳ２０３）。

[0064] 分析部１５３は、文章に含まれる複数の文のうち、ｎ番目の文の文ベクトルＳＶｎを機械学習モデル５０に入力し、ｎ＋１番目の文の文ベクトルＳＶｎ＋１’を予測する（ステップＳ２０４）。

[0065] 分析部１５３は、文章に含まれる複数の文のうち、ｎ＋１番目の文の文ベクトルＳＶｎ＋１と、予測した文の文ベクトルＳＶｎ＋１’とのコサイン類似度を算出する（ステップＳ２０５）。

[0066] 分析部１５３は、コサイン類似度が閾値以上である場合には（ステップＳ２０６，Ｙｅｓ）、ステップＳ２１０に移行する。

[0067] 一方、分析部１５３は、コサイン類似度が閾値以上でない場合には（ステップＳ２０６，Ｎｏ）、ｎ＋１番目の文を、文ベクトルが不適正な文として検出する（ステップＳ２０７）。分析部１５３は、予測した文ベクトルＳＶｎ＋１’と、文転置インデックス１４４とを基にして、文ベクトルＳＶｎ＋１’に対応する文をＤＢ１４３から検出する（ステップＳ２０８）。

[0068] 分析部１５３は、文ベクトルが不適正な文と、ＤＢ１４３から検出した文とを、表示部１３０に表示する（ステップＳ２０９）。

[0069] ステップＳ２１０以降の処理について説明する。分析部１５３は、ｎがＬ以上である場合には（ステップＳ２１０，Ｙｅｓ）、処理を終了する。Ｌは、処理対象の文章に含まれる文の数である。分析部１５３は、ｎがＬ以上でない場合には（ステップＳ２１０，Ｎｏ）、ｎに１を加算した値によって、ｎを更新し（ステップＳ２１１）、ステップＳ２０４に移行する。

[0070] 次に、本実施例に係る情報処理装置１００の効果について説明する。情報処理装置１００は、教師データ１４１に含まれる文章の各文のベクトルを順番に機械学習モデル５０に入力することで、ある第１文の文ベクトルが入力された場合に、第１文の次の第２文の文ベクトルを予測する機械学習モデル５０を生成する。情報処理装置１００は、生成した機械学習モデル５０に、処理対象の文章の文の文ベクトルを入力し、次の文の文ベクトルを予測し、予測した文ベクトルを基にして、処理対象の文章から、不適正な文ベクトルを持つ文を検出する。また、その不適正な文から入力誤り等の単語を適正化することができる。

[0071] 情報処理装置１００は、機械学習モデル５０によって予測された次の文の文ベクトルと、処理対象の文章に含まれる文の次の文の文ベクトルとのコサイン類似度を基にして、文ベクトルの不適正な文を検出し、入力誤り等を適正化する。これによって、計算コストを抑えて、文ベクトルの不適正な文を検出し、入力誤り等を適正化することができる。

[0072] 情報処理装置１００は、帰納法または演繹法に基づいて並び順が決定された複数の文のベクトルを順番に機械学習モデルに入力して訓練する。これによって、帰納法または演繹法に基づいた対象文の次の文を予測することができる。

[0073] 情報処理装置１００は、修正対象の文であると判定された場合に、機械学習モデル５０に予測されたベクトルを基にして、修正後の文を検索する。これによって、修正後の文を通知することができる。

[0074] なお、上述した情報処理装置１００の処理内容は一例であり、情報処理装置１００は、その他の処理を実行してもよい。以下では、情報処理装置１００のその他の処理について説明する。

[0075] 図９及び図１０は、情報処理装置のその他の処理を説明するための図である。上記の情報処理装置１００は、機械学習モデル５０に、三段論法に基づく文のベクトルの順番を学習させていたが、文のベクトルの代わりに、タンパク質の配列であり、単語に相当する複数のアミノ酸配列で構成されるタン

パク質一次構造のベクトルの順番を学習させてもよい。以下の説明では、タンパク質の連続アミノ酸配列を「基本構造」と、タンパク質一次構造を「一次構造」と表記する。

[0076] 図９について説明する。学習フェーズにおいて、情報処理装置１００は、教師データ２４１に含まれる複数のタンパク質の配列２０ａ，２０ｂを用いて、機械学習モデル５０の学習を実行する。

[0077] たとえば、配列２０ａには、一次構造「α一次構造」、「β一次構造」、・・・、「γ一次構造」が含まれる。配列２０ｂには、一次構造「Δ一次構造」、「ε一次構造」、・・・、「ζ一次構造」が含まれる。

[0078] 情報処理装置１００は、基本構造と、ベクトルとを対応付けたタンパク質の基本構造のベクトル辞書を用いて、各一次構造のベクトルを特定する。たとえば、複数の基本構造で構成される一次構造「α一次構造」のベクトルを「Ｖ２０－１」、一次構造「β一次構造」のベクトルを「Ｖ２０－２」、一次構造「γ一次構造」のベクトルを「Ｖ２０－３」とする。一次構造「Δ一次構造」のベクトルを「Ｖ２１－１」、一次構造「ε一次構造」のベクトルを「Ｖ２１－２」、一次構造「ζ一次構造」のベクトルを「Ｖ２１－３」とする。各一次構造のベクトルは、その一次構造を構成する複数の基本構造の各基本構造のベクトルをもとに算出される。

[0079] 情報処理装置１００は、タンパク質の配列に含まれる先頭の一次構造のベクトルから順番に、機械学習モデル５０に入力する処理を繰り返し実行する。たとえば、情報処理装置は、ベクトル「Ｖ２０－１」、「Ｖ２０－２」、・・・、「Ｖ２０－３」の順に、機械学習モデル５０にベクトルを入力する。情報処理装置は、ベクトル「Ｖ２１－１」、「Ｖ２１－２」、・・・、「Ｖ２１－３」の順に、機械学習モデル５０にベクトルを入力する。

[0080] 情報処理装置１００が、上記の学習フェーズの処理を実行することで、ある一次構造のベクトルが入力された場合に、ある一次構造の次の一次構造のベクトルを予測する機械学習モデル５０が生成される。

[0081] 図１０について説明する。分析フェーズにおいて、処理対象のタンパク質

の配列を、配列２５とする。配列２５には、先頭から順に、一次構造「α一次構造」、「η一次構造」、・・・、「γ一次構造」が含まれる。

[0082] 情報処理装置１００は、一次構造「α一次構造」のベクトル「Ｖ２０－１」を算出し、算出したベクトル「Ｖ２０－１」を、機械学習モデル５０に入力することで、一次構造「α一次構造」の次の一次構造のベクトルを予測する。図１０に示す例では、機械学習モデル５０によって、一次構造「α一次構造」の次の一次構造のベクトルとして、「Ｖ２０－２」が予測されている。

[0083] 情報処理装置１００は、配列２５に含まれる一次構造であって、一次構造「α一次構造」の次の「η一次構造」のベクトル「Ｖ２２」を算出する。

[0084] 情報処理装置１００は、機械学習モデル５０によって予測された次の一次構造のベクトル「Ｖ２０－２」と、配列２５に含まれる一次構造であって、基本構造「α一次構造」の次の「η一次構造」のベクトル「Ｖ２２」とのコサイン類似度を算出する。

[0085] 情報処理装置は、コサイン類似度が閾値以上の場合、配列２５に含まれる一次構造であって、一次構造「α一次構造」の次の「η一次構造」が正しい一次構造であると判定する。一方、情報処理装置は、コサイン類似度が閾値未満の場合、配列２５に含まれる一次構造であって、一次構造「α一次構造」の次の「η一次構造」を不適正な一次構造であると判定し、一次構造「η一次構造」に含まれる基本構造の突然変異等を適正化する。

[0086] 図９及び図１０に示した処理を、情報処理装置１００が実行することで、タンパク質の配列に含まれる複数の一次構造から、不適正な一次構造ベクトルを持つ一次構造を検出し、突然変異等のある基本構造を適正化することができる。これにより、複数のタンパク質一次構造で構成される受容体に発生した突然変異等（ＳＮＰｓが代表例である）を持つタンパク質一次構造を検出することができる。さらに、受容体を構成する多数のタンパク質一次構造と、受容体に結合する単一または複数のタンパク質一次構造を結合順に機械学習することで、受容体に結合するリガンドのタンパク質一次構造のベクト

ルを予測することができる。これにより、既に、バイオ医薬品として製品化されたリガンドと類似し、優れた薬効を持ち、副反応が抑制された、新しいタンパク質一次構造のベクトルを持つリガンドの改良を支援することができる。

[0087] 次に、上記実施例に示した情報処理装置１００と同様の機能を実現するコンピュータのハードウェア構成の一例について説明する。図１１は、実施例の情報処理装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。

[0088] 図１１に示すように、コンピュータ３００は、各種演算処理を実行するＣＰＵ３０１と、ユーザからのデータの入力を受け付ける入力装置３０２と、ディスプレイ３０３とを有する。また、コンピュータ３００は、有線または無線ネットワークを介して、外部装置等との間でデータの授受を行う通信装置３０４と、インタフェース装置３０５とを有する。また、コンピュータ３００は、各種情報を一時記憶するＲＡＭ３０６と、ハードディスク装置３０７とを有する。そして、各装置３０１～３０７は、バス３０８に接続される。

[0089] ハードディスク装置３０７は、前処理プログラム３０７ａ、学習プログラム３０７ｂ、分析プログラム３０７ｃを有する。また、ＣＰＵ３０１は、各プログラム３０７ａ～３０７ｃを読み出してＲＡＭ３０６に展開する。

[0090] 前処理プログラム３０７ａは、前処理プロセス３０６ａとして機能する。学習プログラム３０７ｂは、学習プロセス３０６ｂとして機能する。分析プログラム３０７ｃは、分析プロセス３０６ｃとして機能する。

[0091] 前処理プロセス３０６ａの処理は、前処理部１５１の処理に対応する。学習プロセス３０６ｂの処理は、学習部１５２の処理に対応する。分析プロセス３０６ｃの処理は、分析部１５３の処理に対応する。

[0092] なお、各プログラム３０７ａ～３０７ｃについては、必ずしも最初からハードディスク装置３０７に記憶させておかなくても良い。例えば、コンピュータ３００に挿入されるフレキシブルディスク（ＦＤ）、ＣＤ－ＲＯＭ、Ｄ

ＶＤ、光磁気ディスク、ＩＣカードなどの「可搬用の物理媒体」に各プログラムを記憶させておく。そして、コンピュータ３００が各プログラム３０７ａ～３０７ｃを読み出して実行するようにしてもよい。

符号の説明

[0093]

５０	機械学習モデル
１００	情報処理装置
１１０	通信部
１２０	入力部
１３０	表示部
１４０	記憶部
１４１	教師データ
１４２	単語ベクトル辞書
１４３	ＤＢ
１４４	文転置インデックス
１５０	制御部
１５１	前処理部
１５２	学習部
１５３	分析部

請求の範囲

[請求項1] 連続する複数の文であって、前後の文に関係性を有する前記複数の文のベクトルをそれぞれ算出し、

前記複数の文のベクトルを順番に機械学習モデルに入力して訓練することで、前記機械学習モデルにある文のベクトルを入力した際に、前記ある文の次に入力される文の文ベクトルを予測する前記機械学習モデルを生成し、

第１文のベクトルと、前記第１文に続く第２文のベクトルとを算出し、

前記第１文のベクトルを前記機械学習モデルに入力することで前記第１文に続くと予測される文のベクトルを算出し、前記第２文のベクトルが適正であるか否かを判定する

処理をコンピュータに実行させることを特徴とする情報処理プログラム。

[請求項2] 前記判定する処理は、前記第１文のベクトルを、前記機械学習モデルに入力することで予測されるベクトルと、前記第２文のベクトルとのコサイン類似度を基にして、前記第２文のベクトルが適正であるか否かを判定することを特徴とする請求項１に記載の情報処理プログラム。

[請求項3] 前記連続する複数の文は、帰納法または演繹法に基づいて並び順が決定された複数の文であり、前記機械学習モデルを生成する処理は、帰納法または演繹法に基づいて並び順が決定された複数の文のベクトルを順番に機械学習モデルに入力して訓練することを特徴とする請求項２に記載の情報処理プログラム。

[請求項4] 前記第２文のベクトルが不適正であると判定された場合に、前記第１文のベクトルを前記機械学習モデルに入力することで前記第１文に続くと予測される文のベクトルを算出し、算出したベクトルと類似した文を検索し適正な文の候補として提示するために、算出した前記第

１文に続くと予測される文のベクトルを基にして、適正な文を推薦する処理を更にコンピュータに実行させることを特徴とする請求項１に記載の情報処理プログラム。

[請求項5] 連続する複数の文であって、前後の文に関係性を有する前記複数の文のベクトルをそれぞれ算出し、

前記複数の文のベクトルを順番に機械学習モデルに入力して訓練することで、前記機械学習モデルにある文のベクトルを入力した際に、前記ある文の次に入力される文の文ベクトルを予測する前記機械学習モデルを生成し、

第１文のベクトルと、前記第１文に続く第２文のベクトルとを算出し、

前記第１文のベクトルを前記機械学習モデルに入力することで前記第１文に続くと予測される文のベクトルを算出し、前記第２文のベクトルが適正であるか否かを判定する

処理をコンピュータが実行することを特徴とする情報処理方法。

[請求項6] 前記判定する処理は、前記第１文のベクトルを、前記機械学習モデルに入力することで予測されるベクトルと、前記第２文のベクトルとのコサイン類似度を基にして、前記第２文のベクトルが適正であるか否かを判定することを特徴とする請求項５に記載の情報処理方法。

[請求項7] 前記連続する複数の文は、帰納法または演繹法に基づいて並び順が決定された複数の文であり、前記機械学習モデルを生成する処理は、帰納法または演繹法に基づいて並び順が決定された複数の文のベクトルを順番に機械学習モデルに入力して訓練することを特徴とする請求項６に記載の情報処理方法。

[請求項8] 前記第２文のベクトルが不適正であると判定された場合に、前記第１文のベクトルを前記機械学習モデルに入力することで前記第１文に続くと予測される文のベクトルを算出し、算出したベクトルと類似した文を検索し適正な文の候補として提示するために、算出した前記第

１文に続くと予測される文のベクトルを基にして、適正な文を推薦する処理を更にコンピュータに実行させることを特徴とする請求項６に記載の情報処理方法。

[請求項9] 連続する複数の文であって、前後の文に関係性を有する前記複数の文のベクトルをそれぞれ算出し、

前記複数の文のベクトルを順番に機械学習モデルに入力して訓練することで、前記機械学習モデルにある文のベクトルを入力した際に、前記ある文の次に入力される文の文ベクトルを予測する前記機械学習モデルを生成し、

第１文のベクトルと、前記第１文に続く第２文のベクトルとを算出し、

前記第１文のベクトルを前記機械学習モデルに入力することで前記第１文に続くと予測される文のベクトルを算出し、前記第２文のベクトルが適正であるか否かを判定する

処理を実行する制御部を有することを特徴とする情報処理装置。

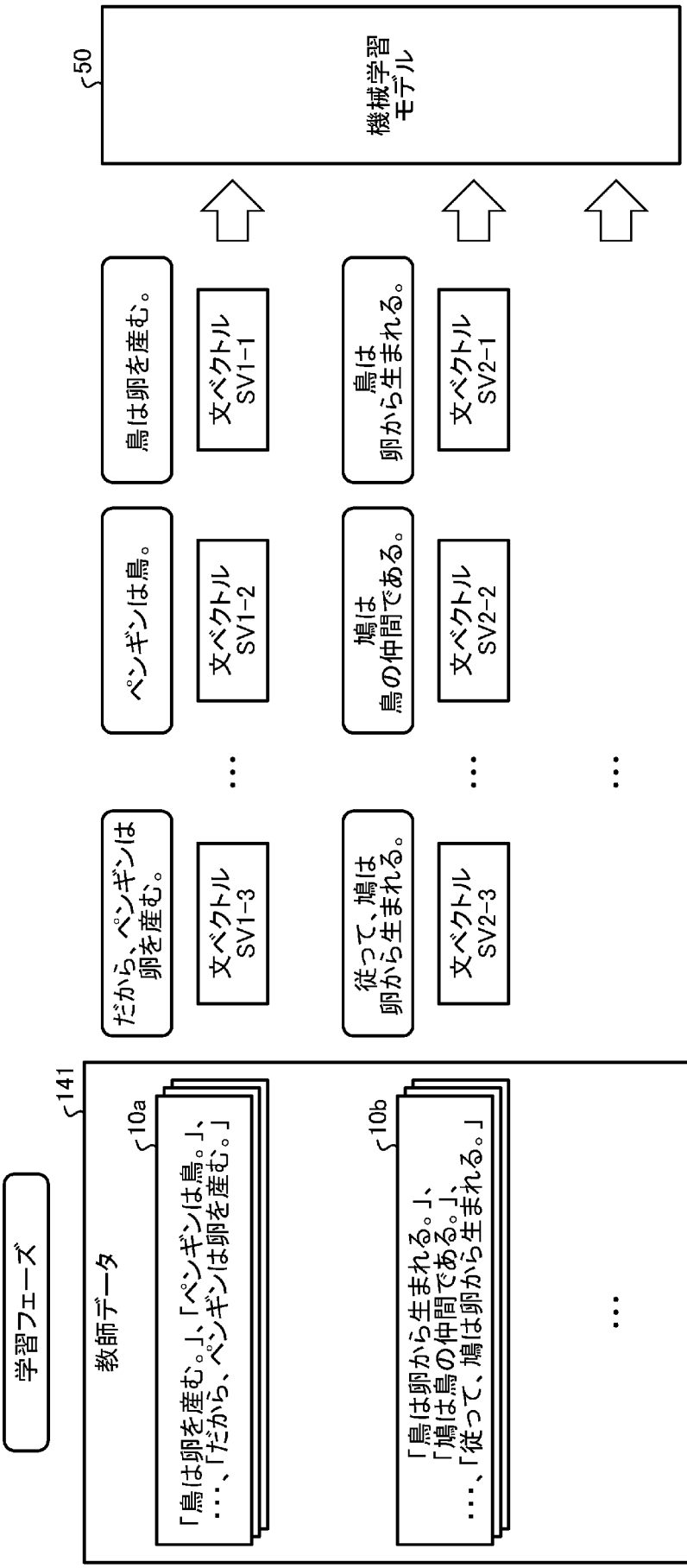
[請求項10] 前記制御部が実行する前記判定する処理は、前記第１文のベクトルを、前記機械学習モデルに入力することで予測されるベクトルと、前記第２文のベクトルとのコサイン類似度を基にして、前記第２文のベクトルが適正であるか否かを判定することを特徴とする請求項９に記載の情報処理装置。

[請求項11] 前記連続する複数の文は、帰納法または演繹法に基づいて並び順が決定された複数の文であり、前記制御部が実行する前記機械学習モデルを生成する処理は、帰納法または演繹法に基づいて並び順が決定された複数の文のベクトルを順番に機械学習モデルに入力して訓練することを特徴とする請求項１０に記載の情報処理装置。

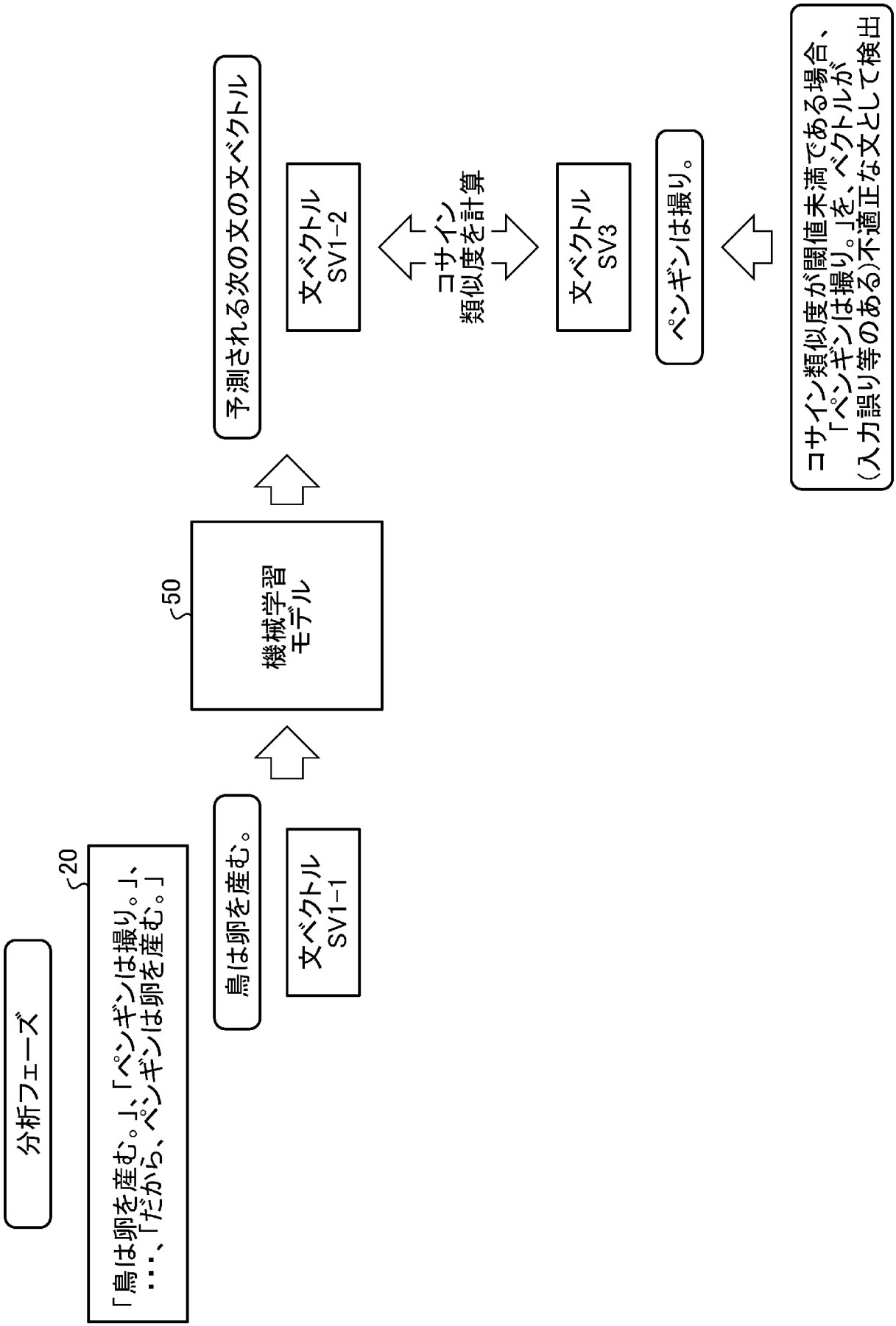
[請求項12] 前記制御部は、前記第２文のベクトルが不適正であると判定された場合に、前記第１文のベクトルを前記機械学習モデルに入力することで前記第１文に続くと予測される文のベクトルを算出し、算出したベ

クトルと類似した文を検索し適正な文の候補として提示するために、算出した前記第１文に続くと予測される文のベクトルを基にして、適正な文を推薦する処理を更にコンピュータに実行させることを特徴とする請求項９に記載の情報処理装置。

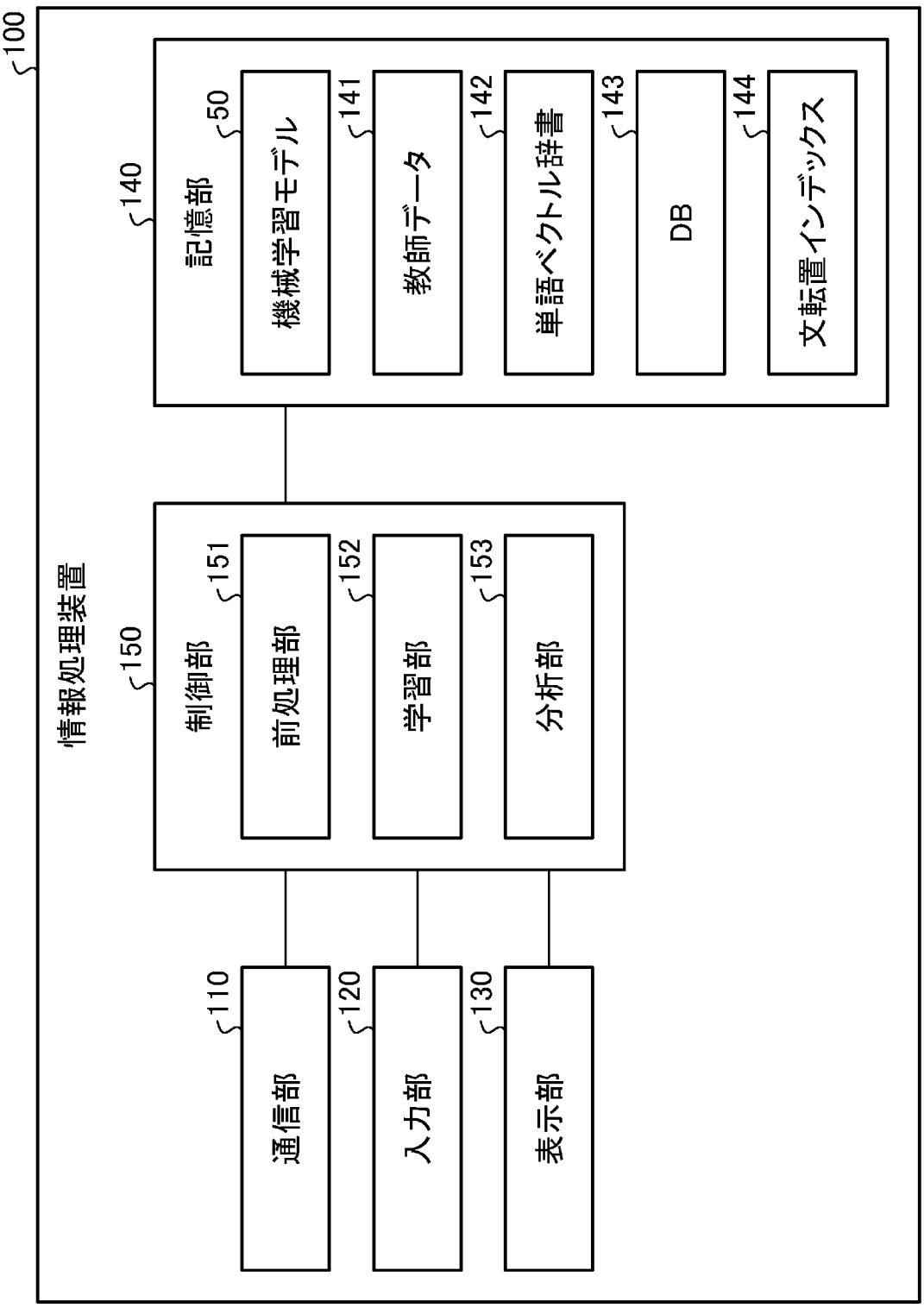
[図1]


学習フェーズ
教師データ
141
10a
「鳥は卵を産む。」、「ペンギンは鳥。」、
…、「だから、ペンギンは卵を産む。」
10b
「鳥は卵から生まれる。」、
「鳩は鳥の仲間である。」、
…、「従って、鳩は卵から生まれる。」
…
だから、ペンギンは
卵を産む。
文ベクトル
SV1-3
ペンギンは鳥。
文ベクトル
SV1-2
鳥は卵を産む。
文ベクトル
SV1-1
…
従って、鳩は
卵から生まれる。
文ベクトル
SV2-3
鳩は
鳥の仲間である。
文ベクトル
SV2-2
鳥は
卵から生まれる。
文ベクトル
SV2-1
…
…
50
機械学習
モデル

[図2]


分析フェーズ
20
「鳥は卵を産む。」、「ペンギンは撮り。」、
…、「だから、ペンギンは卵を産む。」
鳥は卵を産む。
文ベクトル
SV1-1
50
機械学習
モデル
予測される次の文の文ベクトル
文ベクトル
SV1-2
コサイン
類似度を計算
文ベクトル
SV3
ペンギンは撮り。
コサイン類似度が閾値未満である場合、
「ペンギンは撮り。」を、ベクトルが
(入力誤り等のある)不適正な文として検出

[図3]

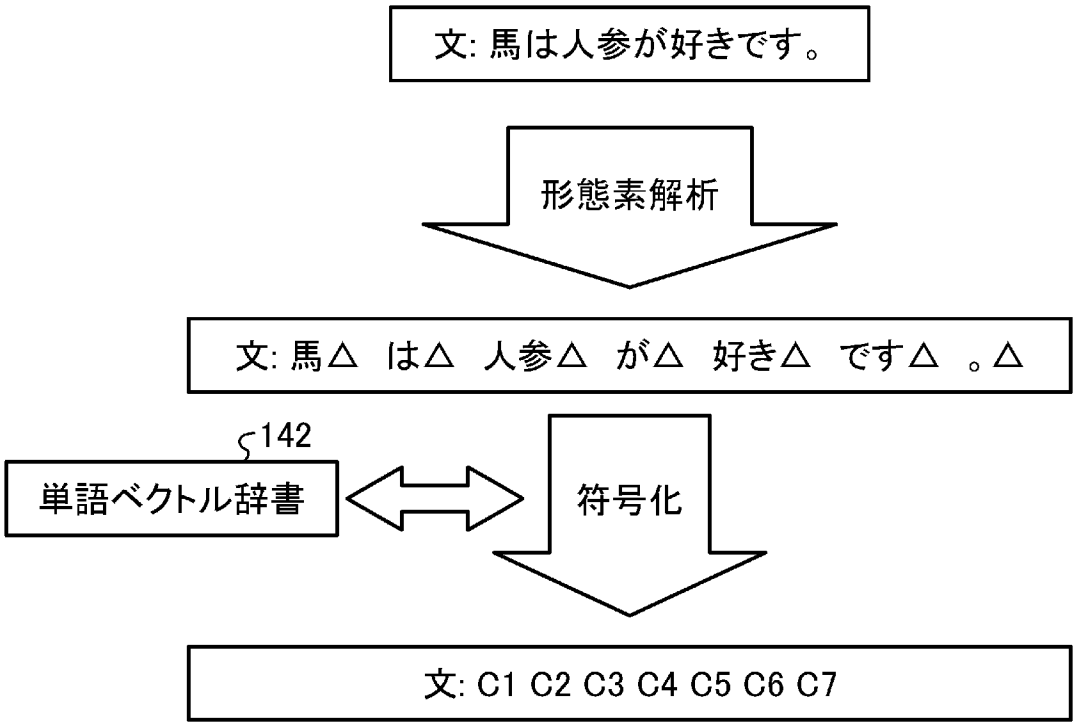

100
情報処理装置
140
記憶部
50
機械学習モデル
141
教師データ
142
単語ベクトル辞書
143
DB
144
文転置インデックス
150
制御部
151
前処理部
152
学習部
153
分析部
110
通信部
120
入力部
130
表示部

[図4]

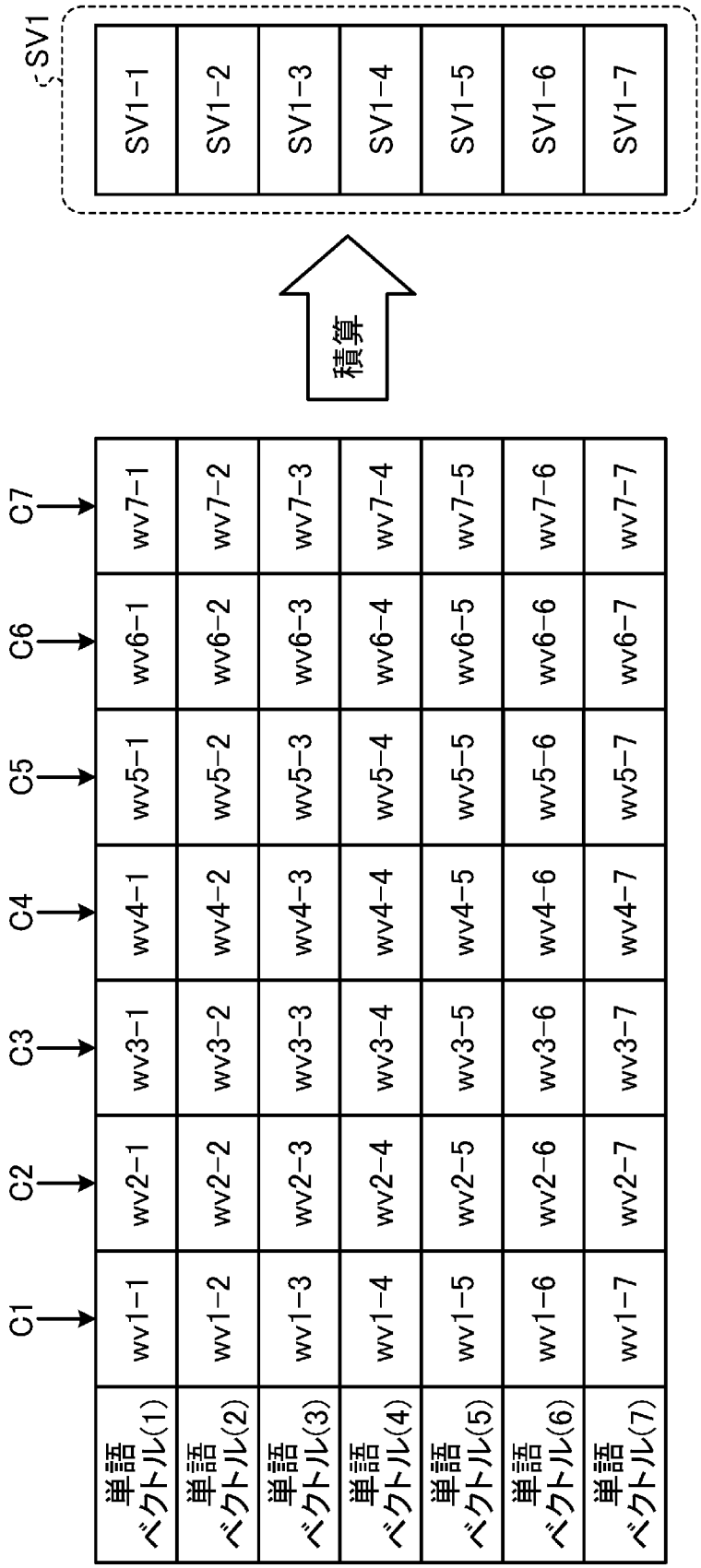
142

符号	単語	単語ベクトル(1)	単語ベクトル(2)	単語ベクトル(3)	単語ベクトル(4)	単語ベクトル(5)	単語ベクトル(6)	単語ベクトル(7)
C1	馬	wv1-1	wv1-2	wv1-3	wv1-4	wv1-5	wv1-6	wv1-7
C2	は	wv2-1	wv2-2	wv2-3	wv2-4	wv2-5	wv2-6	wv2-7
C3	人参	wv3-1	wv3-2	wv3-3	wv3-4	wv3-5	wv3-6	wv3-7
C4	が	wv4-1	wv4-2	wv4-3	wv4-4	wv4-5	wv4-6	wv4-7
C5	好き	wv5-1	wv5-2	wv5-3	wv5-4	wv5-5	wv5-6	wv5-7
C6	です	wv6-1	wv6-2	wv6-3	wv6-4	wv6-5	wv6-6	wv6-7
C7	。	wv7-1	wv7-2	wv7-3	wv7-4	wv7-5	wv7-6	wv7-7
C8	の	wv8-1	wv8-2	wv8-3	wv8-4	wv8-5	wv8-6	wv8-7
C9	好物	wv9-1	wv9-2	wv9-3	wv9-4	wv9-5	wv9-6	wv9-7
C10	今日	wv10-1	wv10-2	wv10-3	wv10-4	wv10-5	wv10-6	wv10-7
C11	天気	wv11-1	wv11-2	wv11-3	wv11-4	wv11-5	wv11-6	wv11-7
C12	快晴	wv12-1	wv12-2	wv12-3	wv12-4	wv12-5	wv12-6	wv12-7
…	…	…	…	…	…	…	…	…

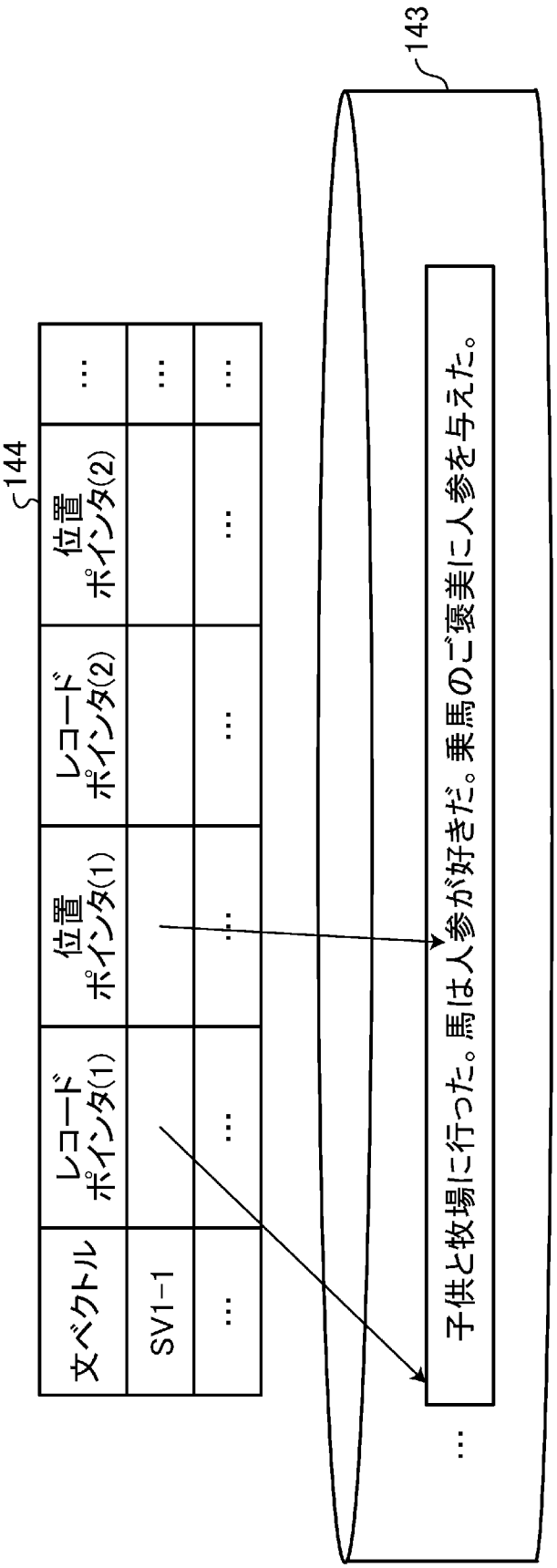
[図5A]


文: 馬は人参が好きです。
形態素解析
文: 馬△ は△ 人参△ が△ 好き△ です△ 。△
142
単語ベクトル辞書
符号化
文: C1 C2 C3 C4 C5 C6 C7

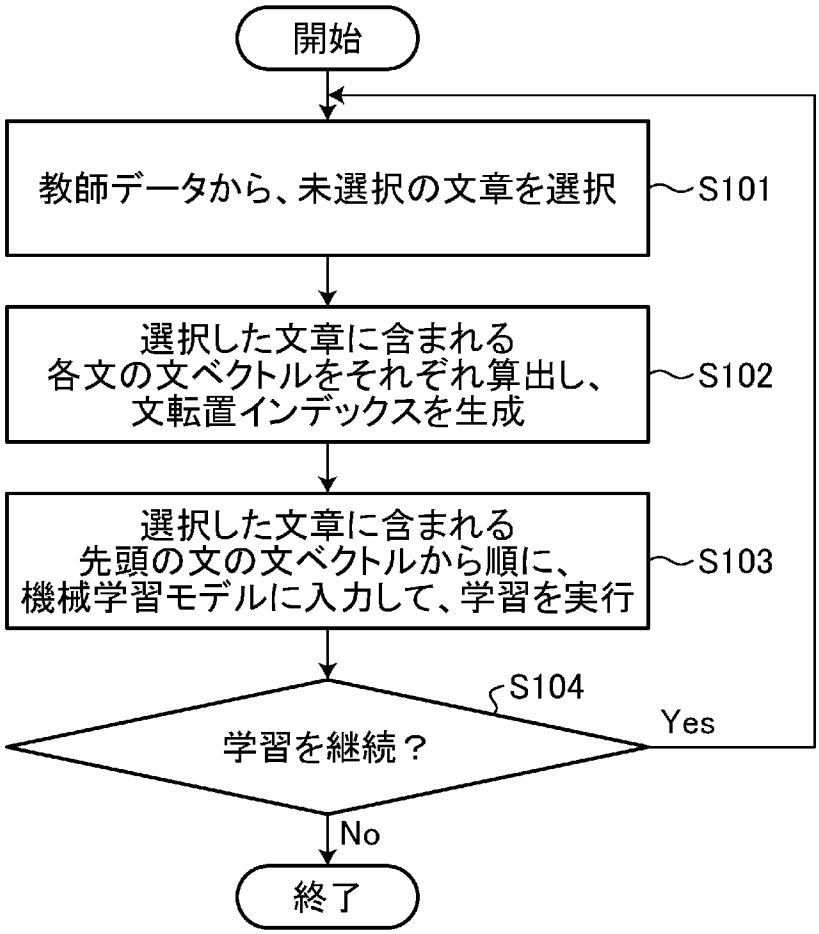
[図5B]


SV1
SV1−1
SV1−2
SV1−3
SV1−4
SV1−5
SV1−6
SV1−7
積算
C1
C2
C3
C4
C5
C6
C7
単語ベクトル(1)
wv1−1
wv2−1
wv3−1
wv4−1
wv5−1
wv6−1
wv7−1
単語ベクトル(2)
wv1−2
wv2−2
wv3−2
wv4−2
wv5−2
wv6−2
wv7−2
単語ベクトル(3)
wv1−3
wv2−3
wv3−3
wv4−3
wv5−3
wv6−3
wv7−3
単語ベクトル(4)
wv1−4
wv2−4
wv3−4
wv4−4
wv5−4
wv6−4
wv7−4
単語ベクトル(5)
wv1−5
wv2−5
wv3−5
wv4−5
wv5−5
wv6−5
wv7−5
単語ベクトル(6)
wv1−6
wv2−6
wv3−6
wv4−6
wv5−6
wv6−6
wv7−6
単語ベクトル(7)
wv1−7
wv2−7
wv3−7
wv4−7
wv5−7
wv6−7
wv7−7

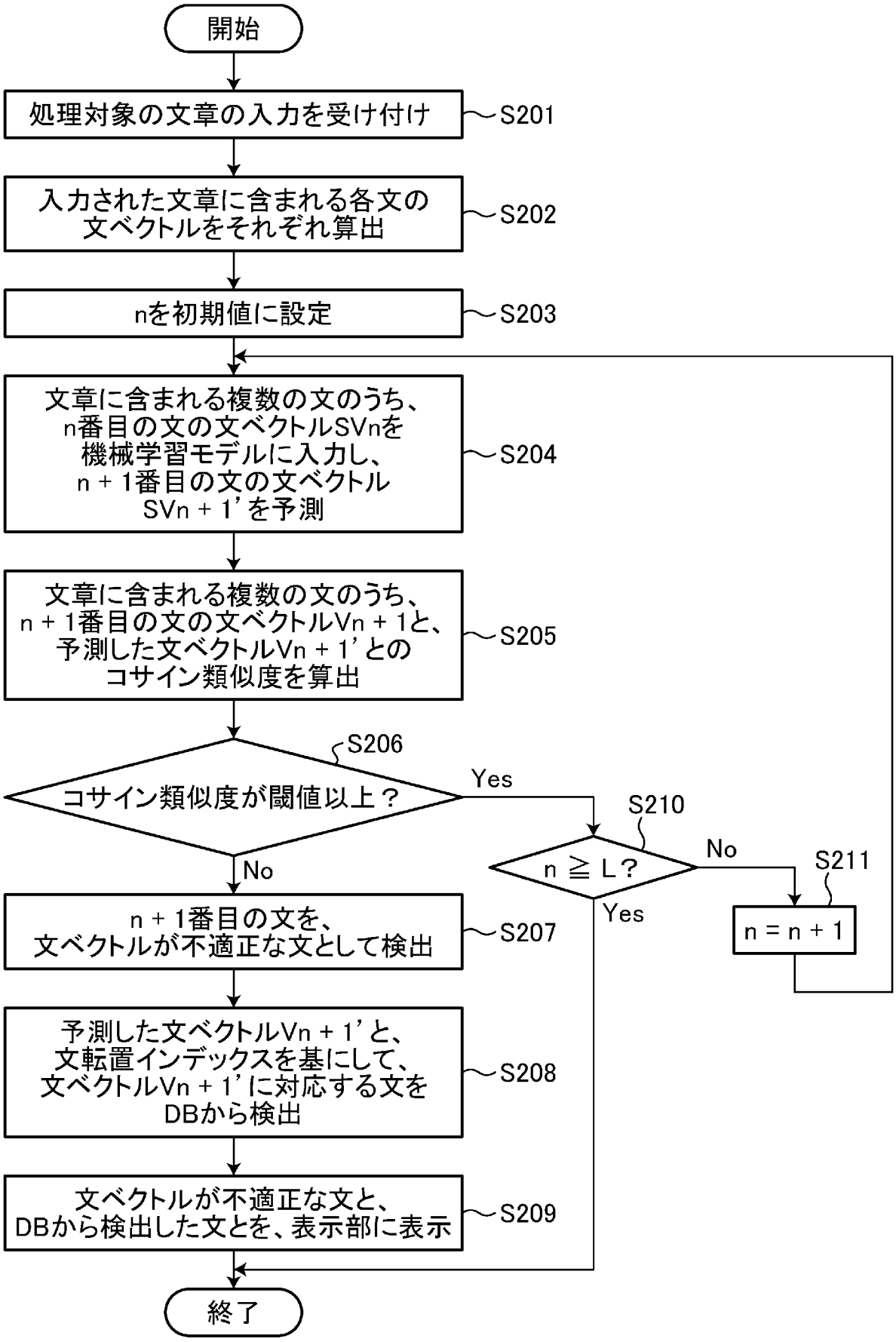
[図6]


143
144
文ベクトル
レコードポインタ(1)
位置ポインタ(1)
レコードポインタ(2)
位置ポインタ(2)
SV1-1
子供と牧場に行った。馬は人参が好きだ。乗馬のご褒美に人参を与えた。

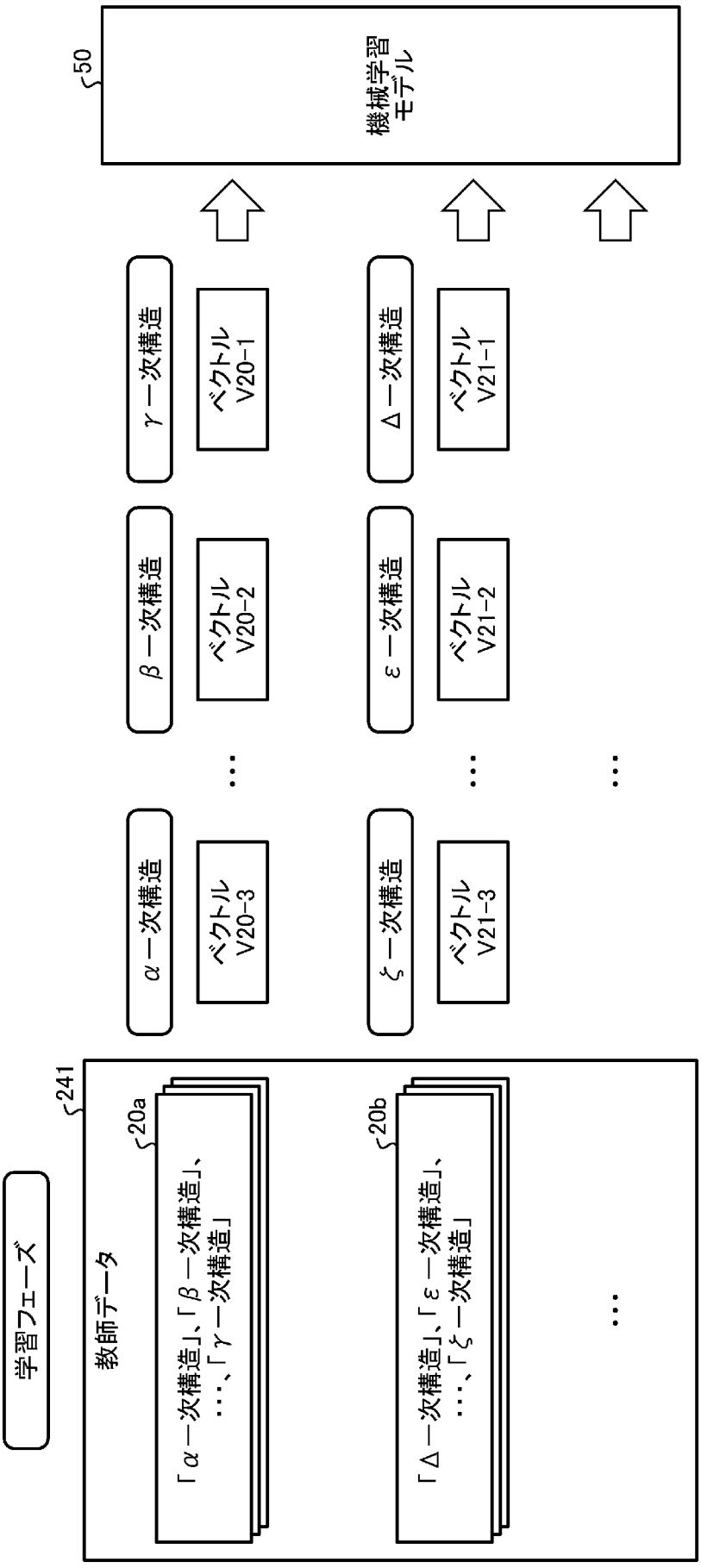
[図7]


開始
教師データから、未選択の文章を選択
S101
選択した文章に含まれる
各文の文ベクトルをそれぞれ算出し、
文転置インデックスを生成
S102
選択した文章に含まれる
先頭の文の文ベクトルから順に、
機械学習モデルに入力して、学習を実行
S103
S104
学習を継続？
Yes
No
終了

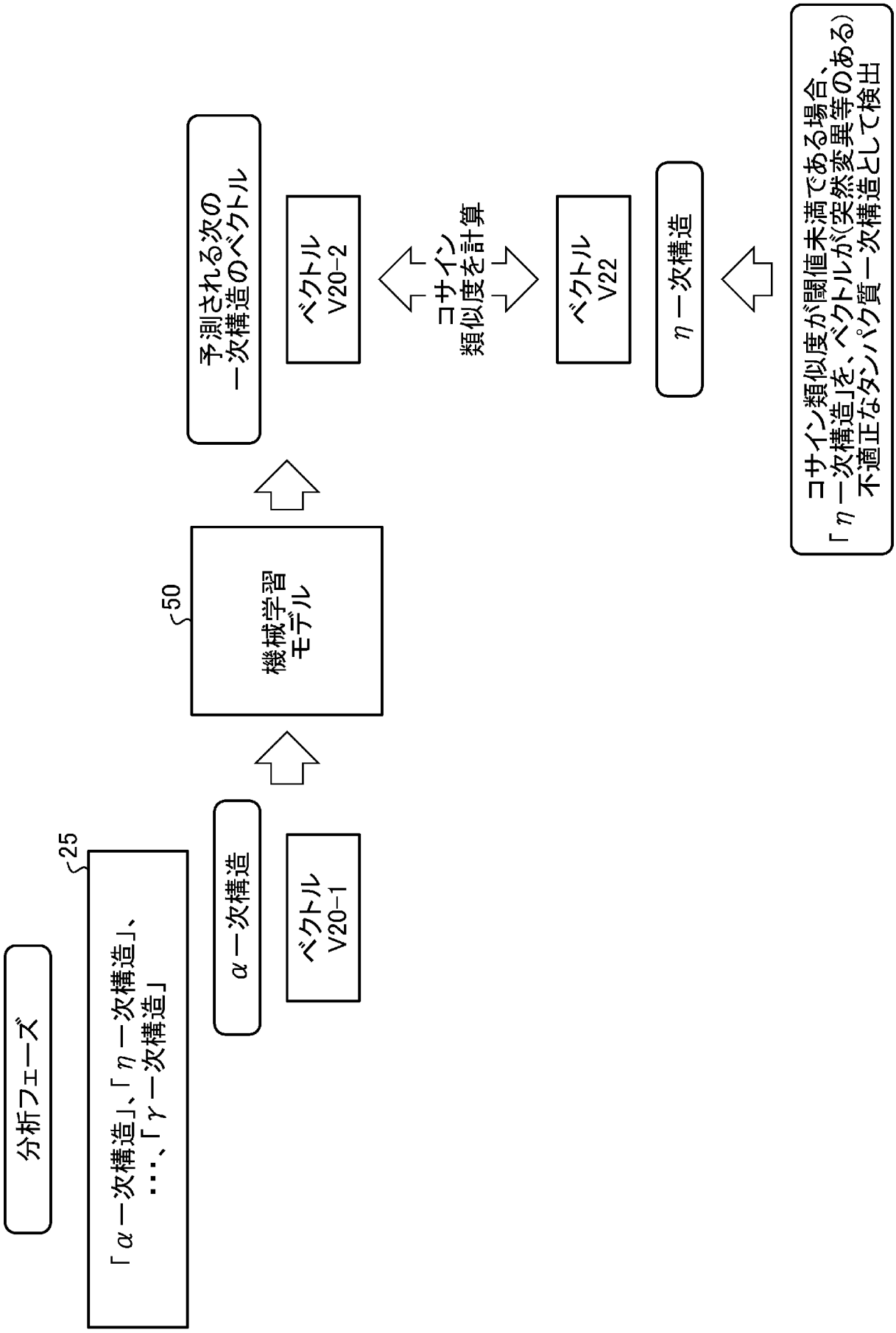
[図8]


開始
処理対象の文章の入力を受け付け
S201
入力された文章に含まれる各文の文ベクトルをそれぞれ算出
S202
nを初期値に設定
S203
文章に含まれる複数の文のうち、n番目の文の文ベクトルSVnを機械学習モデルに入力し、n + 1番目の文の文ベクトルSVn + 1'を予測
S204
文章に含まれる複数の文のうち、n + 1番目の文の文ベクトルVn + 1と、予測した文ベクトルVn + 1'とのコサイン類似度を算出
S205
S206
コサイン類似度が閾値以上？
Yes
No
S210
n ≧ L？
No
Yes
S211
n = n + 1
n + 1番目の文を、文ベクトルが不適正な文として検出
S207
予測した文ベクトルVn + 1'と、文転置インデックスを基にして、文ベクトルVn + 1'に対応する文をDBから検出
S208
文ベクトルが不適正な文と、DBから検出した文とを、表示部に表示
S209
終了

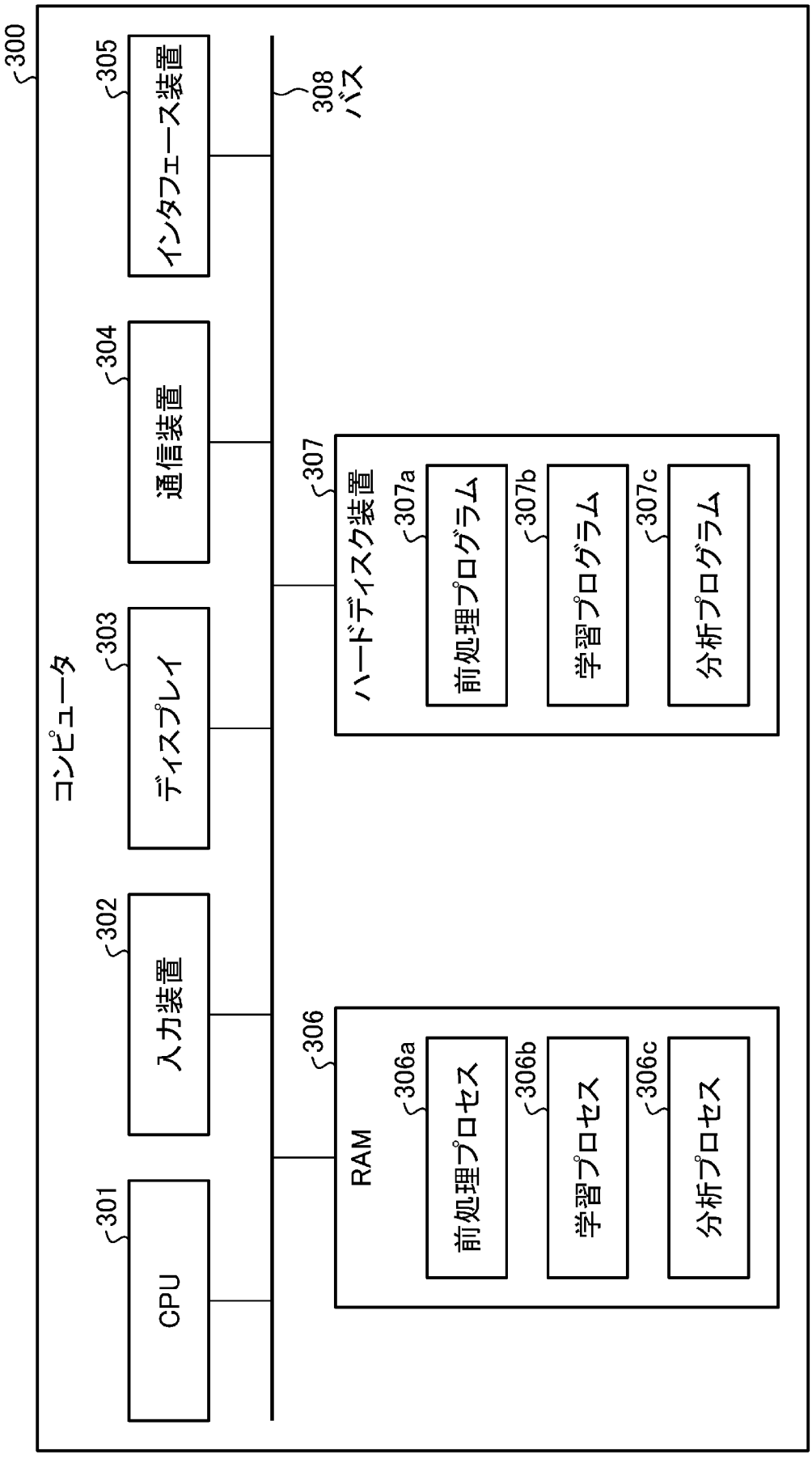
[図9]


学習フェーズ
教師データ
241
20a
「α－次構造」、「β－次構造」、
…、「γ－次構造」
20b
「Δ－次構造」、「ε－次構造」、
…、「ζ－次構造」
α－次構造
ベクトル
V20-3
β－次構造
ベクトル
V20-2
γ－次構造
ベクトル
V20-1
ζ－次構造
ベクトル
V21-3
ε－次構造
ベクトル
V21-2
Δ－次構造
ベクトル
V21-1
50
機械学習
モデル

[図10]


分析フェーズ
25
「α一次構造」、「η一次構造」、
・・・、「γ一次構造」
α一次構造
ベクトル
V20-1
50
機械学習
モデル
予測される次の
一次構造のベクトル
ベクトル
V20-2
コサイン
類似度を計算
ベクトル
V22
η一次構造
コサイン類似度が閾値未満である場合、
「η一次構造」を、ベクトルが(突然変異等のある)
不適正なタンパク質一次構造として検出

[図11]


300
コンピュータ
301
CPU
302
入力装置
303
ディスプレイ
304
通信装置
305
インタフェース装置
306
RAM
306a
前処理プロセス
306b
学習プロセス
306c
分析プロセス
307
ハードディスク装置
307a
前処理プログラム
307b
学習プログラム
307c
分析プログラム
308
バス

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2022/022525

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/166(2020.01)i; ***G06F 40/216***(2020.01)i; ***G06F 40/232***(2020.01)i; ***G06F 40/253***(2020.01)i; ***G06F 40/44***(2020.01)i
FI: G06F40/166; G06F40/216; G06F40/232; G06F40/253; G06F40/44

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)

G06F40/00-40/58

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2022
Registered utility model specifications of Japan 1996-2022
Published registered utility model applications of Japan 1994-2022

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	CN 111539199 A (CHINA MOBILE (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 14 August 2020 (2020-08-14) entire text, all drawings	1-12
A	JP 2019-016140 A (ASAHI SHIMBUN PUBLISHING) 31 January 2019 (2019-01-31) entire text, all drawings	1-12
A	WO 2021/124490 A1 (FUJITSU LIMITED) 24 June 2021 (2021-06-24) entire text, all drawings	1-12
A	JP 2021-089696 A (IB RES KK) 10 June 2021 (2021-06-10) entire text, all drawings	1-12

☐ 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
11 July 2022	**02 August 2022**
Name and mailing address of the ISA/JP **Japan Patent Office (ISA/JP)** **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2022/022525

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	111539199	A	14 August 2020	(Family: none)	
JP	2019-016140	A	31 January 2019	(Family: none)	
WO	2021/124490	A1	24 June 2021	(Family: none)	
JP	2021-089696	A	10 June 2021	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2022/022525

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06F 40/166(2020.01)i; G06F 40/216(2020.01)i; G06F 40/232(2020.01)i; G06F 40/253(2020.01)i; G06F 40/44(2020.01)i
FI: G06F40/166; G06F40/216; G06F40/232; G06F40/253; G06F40/44

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））

G06F40/00-40/58

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	１９７１－２０２２年
日本国実用新案登録公報	１９９６－２０２２年
日本国登録実用新案公報	１９９４－２０２２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	CN 111539199 A (CHINA MOBILE (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) 14.08.2020 (2020-08-14) 全文，全図	1-12
A	JP 2019-016140 A (株式会社朝日新聞社) 31.01.2019 (2019-01-31) 全文，全図	1-12
A	WO 2021/124490 A1 (富士通株式会社) 24.06.2021 (2021-06-24) 全文，全図	1-12
A	JP 2021-089696 A (アイビーリサーチ株式会社) 10.06.2021 (2021-06-10) 全文，全図	1-12

☐ 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抵触するものではなく、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
11.07.2022	02.08.2022

名称及びあて先
日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）
木村　大吾 5N 4807
電話番号 03-3581-1101 内線 3586

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2022/022525

引用文献	公表日	パテントファミリー文献	公表日
CN 111539199 A	14.08.2020	（ファミリーなし）	
JP 2019-016140 A	31.01.2019	（ファミリーなし）	
WO 2021/124490 A1	24.06.2021	（ファミリーなし）	
JP 2021-089696 A	10.06.2021	（ファミリーなし）	