

(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
2022年1月20日(20.01.2022)



(10) 国際公開番号

WO 2022/013954 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号: PCT/JP2020/027411

(22) 国際出願日: 2020年7月14日(14.07.2020)

(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: 野呂 智哉(NORO, Tomoya); 〒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: 特許業務法人酒井国際特許
事務所 (SAKAI INTERNATIONAL PATENT

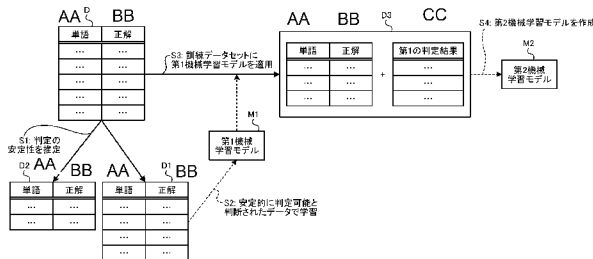
OFFICE); 〒1000013 東京都千代田区霞が
関3丁目8番1号 虎の門三井ビル
ディング Tokyo (JP).

(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, 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, IR, IS, IT, 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.

(54) Title: MACHINE LEARNING PROGRAM, MACHINE LEARNING METHOD, AND INFORMATION PRO-
CESSING DEVICE

(54) 発明の名称: 機械学習プログラム、機械学習方法および情報処理装置

【図1】



M1 ... First machine learning model
M2 ... Second machine learning model
S1 ... Estimate reliability of decision
S2 ... Learn using data for which reliable decision was determined to be possible
S3 ... Apply first machine learning model to training dataset
S4 ... Create second machine learning model
AA ... Word
BB ... Correct
CC ... First decision result

(57) Abstract: A machine learning program according to one embodiment has a computer perform a selection process, a process for generating a first machine learning model, and a process for generating a second training dataset. In the selection process, a plurality of items of data are selected from among a first training dataset on the basis of a frequency of occurrence of identical items of data to which an identical label contained in the first training dataset is attached. In the process for generating the first machine learning model, the first machine learning model is generated by means of machine learning using the selected plurality of items of data. In the process for generating the second training dataset, the second training dataset is generated by combining the first training dataset with the results that the first machine learning model outputs when data contained in the first training dataset is input.

(57) 要約: 実施形態の機械学習プログラムは、選択する処理と、第1の機械学習モデルを生成する処理と、第2の訓練データ群を生成する処理とをコンピュータに実行させる。選択する処理は、第1の訓練データ群に含まれる同一のラベルが付された同一のデータの出現頻度に基づいて、第1の訓練データ群から複数のデータを選択する。第1の機械学習モデルを生成する処理は、選択した複数のデータを用いた機械学習により第1の機械学習モデルを生成する。第2の訓練データ群を生成する処理は、第1の訓練データ群と、第1の訓練データ群に含まれるデータを入力した場合に第1の機械学習モデルが出力する結果とを合わせた第2の訓練データ群を生成する。

[続葉有]

(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, 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))

明 細 書

発明の名称：

機械学習プログラム、機械学習方法および情報処理装置

技術分野

[0001] 本発明の実施形態は、機械学習技術に関する。

背景技術

[0002] 従来、入力データが与えられたとき、それが予め定義されたカテゴリ集合の中のどのカテゴリに属するかを判定する分類タスク（例えば、文章に含まれる各単語の品詞推定、固有表現抽出、語義判定等）を機械学習により生成した機械学習モデルを用いて解く手法が知られている。

[0003] また、訓練データに対する第1機械学習モデルによる出力結果を、第2機械学習モデルへの入力として使用して機械学習を実行するスタッキングと呼ばれる機械学習の手法がある。一般的に、アンサンブル学習の1手法であるスタッキングを用いて積み重ねられた複数の機械学習モデルの推論精度は、単一の機械学習モデルの推論精度よりも良いことが知られている。

[0004] このスタッキングにおいては、例えば、第1機械学習モデルの判定結果の誤りを修正するように第2機械学習モデルの機械学習が実行されても良い。このような第2機械学習モデルを生成するための訓練データを生成する従来技術としては、訓練データをk個のサブセットに分割し、k-1個のサブセットで生成した第1機械学習モデルを用いて残り1個のサブセットに判定結果を追加する。次いで、判定結果を追加する作業を判定結果の追加対象のサブセットを入れ替えながらk回繰り返すことで、第2機械学習モデルの訓練データを生成する手法が知られている。

先行技術文献

非特許文献

[0005] 非特許文献1：Wu et al. N-fold Templated Piped Correction, Proc. of IJCNLP 2004

発明の概要

発明が解決しようとする課題

[0006] しかしながら、上記の従来技術では、分割した k 個のサブセットを入れ替えながら k 回繰り返すことで第 1 機械学習モデルを k 個作らなければならず、効率よく機械学習を行うことが困難であるという問題がある。

[0007] 1 つの側面では、効率的な機械学習を実行できる機械学習プログラム、機械学習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

[0008] 1 つの案では、機械学習プログラムは、選択する処理と、第 1 の機械学習モデルを生成する処理と、第 2 の訓練データ群を生成する処理とをコンピュータに実行させる。選択する処理は、第 1 の訓練データ群に含まれる同一のラベルが付された同一のデータの出現頻度に基づいて、第 1 の訓練データ群から複数のデータを選択する。第 1 の機械学習モデルを生成する処理は、選択した複数のデータを用いた機械学習により第 1 の機械学習モデルを生成する。第 2 の訓練データ群を生成する処理は、第 1 の訓練データ群と、第 1 の訓練データ群に含まれるデータを入力した場合に第 1 の機械学習モデルが出力する結果とを合わせた第 2 の訓練データ群を生成する。

発明の効果

[0009] 効率的な機械学習が実行される。

図面の簡単な説明

[0010] [図1]図 1 は、実施形態の概要を説明する説明図である。

[図2]図 2 は、従来例を説明する説明図である。

[図3]図 3 は、ノイズを加える場合の実施形態の概要を説明する説明図である。

[図4]図 4 は、実施形態にかかる情報処理装置の機能構成例を示すブロック図である。

[図5A]図 5 A は、訓練データセットの一例を説明する説明図である。

[図5B]図5 Bは、出現頻度データの一例を説明する説明図である。

[図5C]図5 Cは、エントロピーデータの一例を説明する説明図である。

[図5D]図5 Dは、自己情報量データの一例を説明する説明図である。

[図5E]図5 Eは、スコアデータの一例を説明する説明図である。

[図6A]図6 Aは、訓練データ安定性判定処理の一例を示すフローチャートである。

[図6B]図6 Bは、訓練データ安定性判定処理の一例を示すフローチャートである。

[図7]図7は、訓練データ安定性判定処理の変形例を示すフローチャートである。

[図8]図8は、判定手法選択処理の概要を説明する説明図である。

[図9]図9は、判定手法選択処理の一例を示すフローチャートである。

[図10A]図10 Aは、判定結果の追加に関する処理例を示すフローチャートである。

[図10B]図10 Bは、結果データの一例を説明する説明図である。

[図11A]図11 Aは、判定結果の追加に関する処理例を示すフローチャートである。

[図11B]図11 Bは、結果データの一例を説明する説明図である。

[図12A]図12 Aは、判定結果の追加に関する処理例を示すフローチャートである。

[図12B]図12 Bは、結果データの一例を説明する説明図である。

[図13]図13は、コンピュータ構成の一例を示すブロック図である。

発明を実施するための形態

[0011] 以下、図面を参照して、実施形態にかかる機械学習プログラム、機械学習方法および情報処理装置を説明する。実施形態において同一の機能を有する構成には同一の符号を付し、重複する説明は省略する。なお、以下の実施形態で説明する機械学習プログラム、機械学習方法および情報処理装置は、一例を示すに過ぎず、実施形態を限定するものではない。また、以下の各実施

形態は、矛盾しない範囲内で適宜組みあわせてもよい。

[0012] 図1は、実施形態の概要を説明する説明図である。図1に示すように、本実施形態では、文章中の各単語（部分文字列）に対して固有表現を示す「固有表現ラベル」を割り当てる分類タスクを、スタッキングの手法を用いて機械学習された第1機械学習モデルM1および第2機械学習モデルM2により解く際の、第1機械学習モデルM1および第2機械学習モデルM2が、訓練データセットDを用いた機械学習で生成される。

[0013] なお、分類タスクは、上記の例に限定するものではなく、単語の品詞推定や語義判定でもよい。また、分類タスクは、機械学習によって生成された機械学習モデルを用いて解くものであればよく、文書中の単語に関する分類以外にも、血圧、心拍数等の生体データより身体異常の有無を分類するものや、各教科の評価および中間・期末試験の得点などの成績データより対象者（受験者）の可否を分類するものであってもよい。よって、機械学習モデルの生成に用いる訓練データセットに含まれるデータ（以下、事例）については、分類タスクに応じて学習対象とする事例であればよい。例えば、身体異常の有無を分類する機械学習モデルを生成する場合は、学習対象ごとの生体データと、その生体データに対する正解（身体異常の有無）などが各事例に含まれることとなる。

[0014] 訓練データセットDには、各事例（例えば文章中の各単語）について、その事例における正解の「固有表現ラベル」を示す正解ラベルが付与されている。本実施形態では、訓練データセットDを用いて教師あり学習を行うことで、勾配ブースティング木（GBM）、ニューラルネットワークなどの第1機械学習モデルM1および第2機械学習モデルM2が生成される。

[0015] 具体的には、本実施形態では、訓練データセットDに含まれる各事例について、全事例の中で同じ正解ラベルが付与された同じ内容の事例が出現する頻度（出現頻度）をもとに、訓練データセットDを用いた機械学習モデルによる判定の安定性を推定する（S1）。頻度は、絶対頻度、相対頻度、累積頻度のいずれかであっても良い。また、出現頻度に基づいて算出される割合

をもとに、各事例の安定性が推定されても良い。また、「同じ内容の事例」は、同一のラベルが付された同一のデータであり、本実施形態では、このようなデータ毎の出現頻度をもとに安定性を推定するものとする。

[0016] 訓練データセットDに含まれる各事例についての、訓練データセットDを用いた機械学習モデルによる判定の安定性とは、各事例が訓練データセットDを用いた機械学習モデルで安定的に判定可能であるか否かを意味する。例えば、安定的に判定可能な事例では、k分割交差検定においてどのように訓練データセットDを分割・学習して得られた機械学習モデルであっても同じ判定結果が得られると推定される。この安定的に判定可能な事例は、訓練データセットDにおいて同一の正解ラベルが付与された同じ内容の事例が多いものや、分類先カテゴリの曖昧性が低いものが相当することから、同じ正解ラベルが付与された同じ内容の事例の出現頻度をもとに推定可能である。逆に、判定結果が不安定な事例は、k分割交差検定における分割方法によっては異なる判定結果が得られると推定される事例である。この判定結果が不安定な事例は、訓練データセットDにおいて同じ内容の事例が少ないものや、分類先カテゴリの曖昧性が高いものが相当することから、同じ正解ラベルが付与された同じ内容の事例の出現頻度をもとに推定可能である。

[0017] 本実施形態では、S1での推定結果をもとに、訓練データセットDより、安定的に判定可能な事例を選択した訓練データセットD1と、訓練データセットD1以外の訓練データセットD2とに分ける。次いで、本実施形態では、安定的に判定可能と判断されたデータ（訓練データセットD1）で機械学習を行って第1機械学習モデルM1を生成する（S2）。次いで、本実施形態では、訓練データセットDに含まれる各データを第1機械学習モデルM1に入力して、第1機械学習モデルM1が出力した第1の判定結果を訓練データセットDに追加し、訓練データセットD3を生成する（S3）。次いで、本実施形態では、訓練データセットD3を用いた機械学習を行って第2機械学習モデルM2を生成する。

[0018] 出現頻度をもとに安定的に判定可能と推定されたデータを用いた機械学習

により生成した第1機械学習モデルM1では、訓練データセットDを入力データとして推論した場合の第1の判定結果において、不安定になると推定された事例で正解ラベルとは異なる結果（判定結果の誤り）が得られやすくなる。このため、訓練データセットDに第1の判定結果を追加した訓練データセットD3は、第1機械学習モデルM1の判定結果の誤りを修正するように最終判定結果を出力する第2機械学習モデルM2の生成に適したものとなる。

[0019] 図2は、従来例を説明する説明図である。図2に示すように、従来例では、訓練データセットD100をk個のサブセット（ $D100_1 \cdots D100_{k-1}$ 、 $D100_k$ ）に分割し（S101）、k-1個のサブセットで訓練して第1機械学習モデルM101を生成する（S102）。次いで、従来例では、残り1個のサブセットを入力として第1機械学習モデルM101が推論して得られた判定結果をそのサブセットに追加する（S103）。従来例では、このようにして判定結果を追加するデータを入れ替えながらS102、S103をk回繰り返すことで第2機械学習モデルM102の訓練データセットD101を生成する（S104）。次いで、従来例では、作成した訓練データセットD101を用いた機械学習によって第2機械学習モデルM102を作成する（S105）。

[0020] このように、従来例では、第2機械学習モデルM102の訓練データセットD101を生成する過程で、分割したk個のサブセット（ $D100_1 \cdots D100_{k-1}$ 、 $D100_k$ ）を入れ替えながらk回繰り返すことから、第1機械学習モデルM101をk個作成することとなる。これに対し、本実施形態では、例えば、機械学習モデルM1を複数作成することなく、第2機械学習モデルM2の訓練データセットD3を効率的に作成することができ、効率的な機械学習を実行することができる。また、第1機械学習モデルM1および第2機械学習モデルM2それぞれの訓練データセットを事前に用意する単純な手法と比較しても、正解フラグを付与するデータ量が少なく済むことから、効率的に機械学習を実行することができる。

[0021] 図3は、ノイズを加える場合の実施形態の概要を説明する説明図である。

図3に示すように、訓練データセットDに第1機械学習モデルM1を適用した第1の判定結果を訓練データセットDに追加し、第2機械学習モデルM2の訓練データセットD3を生成する際には、ノイズを加えてもよい。具体的には、第1機械学習モデルM1の入力にノイズを加えて得られた第1の判定結果を訓練データセットDに追加してもよい（S5a）。または、訓練データセットDに第1機械学習モデルM1を適用して第1機械学習モデルM1が出力する結果にノイズを加えて訓練データセットDに追加してもよい（S5b）。

[0022] このようにノイズを加えた場合、ノイズを加えても結果が変化しにくい事例は第1機械学習モデルM1で安定的に判定可能であり、ノイズを加えると結果が変化しやすい事例は第1機械学習モデルM1では判定結果が不安定となる。したがって、第1機械学習モデルM1による第1の判定結果を訓練データセットDに追加することで、第1機械学習モデルM1の判定結果の誤りを修正するように第2機械学習モデルM2を生成するための訓練データセットD3を生成することができ、第2機械学習モデルM2による最終判定結果の精度を向上させることができる。

[0023] 図4は、実施形態にかかる情報処理装置の機能構成例を示すブロック図である。図4に示すように、情報処理装置1は、入出力部10と、記憶部20と、制御部30とを有する。例えば、情報処理装置1は、PC（Personal Computer）などを適用できる。

[0024] 入出力部10は、制御部30が各種情報の入出力を行う際の入出力インタフェースを司る。例えば、入出力部10は、情報処理装置1に接続されるキーボードやマイク等の入力装置や液晶ディスプレイ装置などの表示装置との入出力インタフェースを司る。また、入出力部10は、LAN（Local Area Network）等の通信ネットワークを介して接続する外部機器との間でデータ通信を行う通信インタフェースを司る。

[0025] 例えば、情報処理装置1は、入出力部10を介して訓練データセットDな

どの入力を受け、記憶部20に格納する。また、情報処理装置1は、生成した第1機械学習モデルM1および第2機械学習モデルM2に関する第1機械学習モデル情報21および第2機械学習モデル情報22を記憶部20より読み出し、入出力部10を介して外部に出力する。

[0026] 記憶部20は、例えば、RAM (Random Access Memory)、フラッシュメモリ (Flash Memory) などの半導体メモリ素子や、HDD (Hard Disk Drive) などの記憶装置によって実現される。記憶部20は、訓練データセットD、出現頻度データ S_f 、エントロピーデータ S_h 、自己情報量データ S_i 、スコアデータ S_d 、訓練データセットD3、第1機械学習モデル情報21および第2機械学習モデル情報22などを格納する。

[0027] 訓練データセットDは、学習対象とする事例（例えば複数の文章それぞれに含まれる各単語）と、その事例に付与された正解ラベル（例えば「固有表現ラベル」）の組（事例と正解ラベルとのペア）についての複数の訓練データの集合である。なお、訓練データは、1文章単位のデータであって、複数の事例と正解ラベルとのペアを含むものとする。

[0028] 図5Aは、訓練データセットDの一例を説明する説明図である。図5Aに示すように、訓練データセットDには、複数の文章それぞれの訓練データに対応するデータIDごとに、文章に含まれる単語と、その単語に付与された正解ラベル（「固有表現ラベル」）との組、すなわち事例と正解ラベルとのペアが含まれる。

[0029] 「固有表現ラベル」については、「O」、「General」、「Molecular」がある。「O」は、固有表現（一部を含む）ではない単語を意味するラベルである。「General」は、「General」という型の固有表現（一部を含む）の単語を意味するラベルである。「Molecular」は、「Molecular」という型の固有表現（一部を含む）の単語を意味するラベルである。なお、「General」および「Molecular」では、先頭の単語には「Bー」という接頭辞を付与し、2番目以降の単語には「Iー」という接頭辞を付与するものとする。

- [0030] 例えば、図示例における訓練データセットDにおいて、「solvent mixture」という事例は、「General」という型の固有表現が正解となる。また、「n-propyl bromide」という事例は、「Molecular」という型の固有表現が正解となる。
- [0031] 図4に戻り、出現頻度データ S_f は、訓練データセットDに含まれる事例と正解ラベルとのペアの出現頻度を集計したデータである。
- [0032] 図5Bは、出現頻度データ S_f の一例を説明する説明図である。図5Bに示すように、出現頻度データ S_f は、訓練データセットDに含まれる各事例について、正解ラベルごとに集計した出現頻度が含まれる。より具体的には、出現頻度データ S_f は、同じ内容の事例と、同じ正解ラベル毎に集計した出現頻度が含まれる。例えば、「solvent mixture」という事例について、正解ラベルが「General」の出現頻度は3である。同様に、「n-propyl bromide」という事例について、正解ラベルが「Molecular」の出現頻度は5である。また、「water」という事例について、正解ラベルが「Molecular」の出現頻度は2083であり、正解ラベルが「General」の出現頻度は5である。
- [0033] 図4に戻り、エントロピーデータ S_h は、訓練データセットDに含まれる各事例について、訓練データセットDに含まれる事例の総事例数、同じ内容の事例と、同じ正解ラベル毎に集計した出現頻度などをもとに計算した情報理論におけるエントロピーを示す。
- [0034] 図5Cは、エントロピーデータ S_h の一例を説明する説明図である。図5Cに示すように、エントロピーデータ S_h は、「solvent mixture」、「n-propyl bromide」、「water」等の各事例のエントロピーを示す。
- [0035] 図4に戻り、自己情報量データ S_i は、訓練データセットDに含まれる事例の総事例数、同じ内容の事例と、同じ正解ラベル毎の出現頻度などをもとに計算した自己情報量を示す。
- [0036] 図5Dは、自己情報量データ S_i の一例を説明する説明図である。図5Dに

示すように、自己情報量データ S_i は、「solvent mixture」と「General」、「n-propyl bromide」と「Molecular」等の同じ内容の事例と、同じ正解ラベル毎の自己情報量を示す。

[0037] 図4に戻り、スコアデータ S_d は、訓練データセットDに含まれる文章それぞれについて、前述した判定の安定性をスコア化したデータである。

[0038] 図5Eは、スコアデータ S_d の一例を説明する説明図である。図5Eに示すように、スコアデータ S_d は、訓練データセットDに含まれる複数の文章それぞれに対応するデータIDごとの、判定の安定性についてのスコアを示す。

[0039] 図4に戻り、第1機械学習モデル情報21は、教師あり学習を行うことで生成した第1機械学習モデルM1に関する情報である。第2機械学習モデル情報22は、教師あり学習を行うことで生成した第2機械学習モデルM2に関する情報である。この第1機械学習モデル情報21および第2機械学習モデル情報22は、例えば勾配ブースティング木、ニューラルネットワークなどのモデルを構築するためのパラメータ等である。

[0040] 制御部30は、第1機械学習モデル生成部31、訓練データ生成部32および第2機械学習モデル生成部33を有する。制御部30は、CPU (Central Processing Unit) やMPU (Micro Processing Unit) などによって実現できる。また、制御部30は、ASIC (Application Specific Integrated Circuit) やFPGA (Field Programmable Gate Array) などのハードワイヤードロジックによっても実現できる。

[0041] 第1機械学習モデル生成部31は、訓練データセットDを用いて第1機械学習モデルM1を生成する処理部である。具体的には、第1機械学習モデル生成部31は、訓練データセットDに含まれる同じ正解ラベルが付された同じ内容の事例毎の出現頻度に基づいて、訓練データセットDから複数の事例を選択する。これにより、第1機械学習モデル生成部31は、訓練データセットDより安定的に判定可能な事例を選択した訓練データセットD1を得る。次いで、第1機械学習モデル生成部31は、訓練データセットD1に含ま

れる複数の事例を用いた機械学習により第1機械学習モデルM1を生成する。次いで、第1機械学習モデル生成部31は、生成した第1機械学習モデルM1に関する第1機械学習モデル情報21を記憶部20に格納する。

[0042] 訓練データ生成部32は、第2機械学習モデルM2を生成するための訓練データセットD3を生成する処理部である。具体的には、訓練データ生成部32は、第1機械学習モデル情報21をもとに第1機械学習モデルM1を構築する。次いで、訓練データ生成部32は、訓練データセットDに含まれる各事例を構築した第1機械学習モデルM1に入力した場合に第1機械学習モデルM1が出力する結果を訓練データセットDに追加して訓練データセットD3を生成する。

[0043] 第2機械学習モデル生成部33は、訓練データセットD3を用いて第2機械学習モデルM2を生成する処理部である。具体的には、第2機械学習モデル生成部33は、訓練データセットD3に含まれる各事例と、その事例に対する第1機械学習モデルM1の判定結果（第1機械学習モデルM1が出力した結果）とを用いた機械学習により第2機械学習モデルM2を生成する。次いで、第2機械学習モデル生成部33は、生成した第2機械学習モデルM2に関する第2機械学習モデル情報22を記憶部20に格納する。

[0044] ここで、第1機械学習モデル生成部31および訓練データ生成部32の処理の詳細を説明する。まず、第1機械学習モデル生成部31では、訓練データセットDの中の各事例の出現頻度をもとに、各事例の判定結果の安定性を示すスコアを算出して訓練データセットD1を得る訓練データ安定性判定処理を行う（S10）。

[0045] 図6A、図6Bは、訓練データ安定性判定処理の一例を示すフローチャートである。図6Aに示すように、処理が開始されると、第1機械学習モデル生成部31は、事例と正解ラベルとのペアを訓練データセットDから収集してその出現頻度を集計する処理（S20）を行う。

[0046] 具体的には、第1機械学習モデル生成部31は、訓練データセットD中のデータIDの集合を処理用の配列（I）等に格納する（S21）。次いで、

第1機械学習モデル生成部31は、配列(1)内のデータIDが空であるか否かを判定し(S22)、空と判定(S22:Yes)されるまでS23～S25の処理を繰り返す。

[0047] 配列(1)内のデータIDが空でないと判定された場合(S22:No)、第1機械学習モデル生成部31は、配列(1)からデータIDを1つ取得し、処理用の変数(id)に格納する(S23)。このとき、第1機械学習モデル生成部31は、取得したデータIDを配列(1)から消去する。次いで、第1機械学習モデル生成部31は、訓練データセットD中の変数(id)に対応するデータから同じ内容の事例と、同じ正解ラベルのペアを取得し(S24)、取得した数(出現頻度)をもとに出現頻度データ S_f を更新する(S25)。

[0048] 配列(1)内のデータIDが空と判定された場合(S22:Yes)、第1機械学習モデル生成部31は、収集した事例ごとのエントロピーと、同じ内容の事例と、同じ正解ラベル毎の自己情報量を計算する処理(S30)を行う。

[0049] 具体的には、第1機械学習モデル生成部31は、出現頻度データ S_f 中の事例集合を処理用の配列(E)等の格納する(S31)。次いで、第1機械学習モデル生成部31は、配列(E)内の事例が空であるか否かを判定し(S32)、空と判定(S32:Yes)されるまでS33～S35の処理を繰り返す。

[0050] 配列(E)内の事例が空でないと判定された場合(S32:No)、第1機械学習モデル生成部31は、第1機械学習モデル生成部31は、配列(E)から事例を1つ取得し、処理用の変数(ex)に格納する(S33)。このとき、第1機械学習モデル生成部31は、取得した事例を配列(E)から消去する。次いで、第1機械学習モデル生成部31は、訓練データセットD中の変数(ex)に対応する事例を検索し、正解ラベルごとに該当数を集計する(S34)。次いで、第1機械学習モデル生成部31は、S34の集計結果をもとに、処理対象の事例と正解ラベルとのペアについて、公知の情報

理論におけるエントロピーと自己情報量を計算し、計算結果をもとにエントロピーデータ S_h および自己情報量データ S_i を更新する (S 3 5)。

[0051] 配列 (E) 内の事例が空と判定された場合 (S 3 2 : Y e s)、図 6 B に示すように、第 1 機械学習モデル生成部 3 1 は、同じ内容の事例と、同じ正解ラベル毎の、前述した判定の安定性を推定する処理を行う (S 4 0)。

[0052] 具体的には、第 1 機械学習モデル生成部 3 1 は、訓練データセット D 中のデータ I D の集合を処理用の配列 (I) 等に格納する (S 4 1)。次いで、第 1 機械学習モデル生成部 3 1 は、配列 (I) 内のデータ I D が空であるか否かを判定し (S 4 2)、空と判定 (S 4 2 : Y e s) されるまで S 4 3 ~ S 4 6 の処理を繰り返す。

[0053] 配列 (I) 内のデータ I D が空でないと判定された場合 (S 4 2 : N o)、第 1 機械学習モデル生成部 3 1 は、配列 (I) からデータ I D を 1 つ取得し、処理用の変数 (i d) に格納する (S 4 3)。このとき、第 1 機械学習モデル生成部 3 1 は、取得したデータ I D を配列 (I) から消去する。

[0054] 次いで、第 1 機械学習モデル生成部 3 1 は、訓練データセット D 中の変数 (i d) に対応するデータから同じ内容の事例と、同じ正解ラベルのペアを取得する (S 4 4)。すなわち、第 1 機械学習モデル生成部 3 1 は、データ I D の文章に関する同じ内容の事例と、同じ正解ラベル毎のペアを取得する。次いで、第 1 機械学習モデル生成部 3 1 は、取得した同じ内容の事例と、同じ正解ラベル毎のペアの出現頻度データ S_f 、エントロピーデータ S_h 、自己情報量データ S_i をもとに、同じ内容の事例と、同じ正解ラベル毎における前述した判定の安定性について、安定・不安定を判定する (S 4 5)。

[0055] 例えば、第 1 機械学習モデル生成部 3 1 は、出現頻度が閾値 (f) 未満であり、訓練データセット D において稀な事例と正解ラベルとのペアを不安定な事例とする。または、第 1 機械学習モデル生成部 3 1 は、自己情報量が閾値 (i) より大きく、かつ、エントロピーが閾値 (h) 未満の曖昧性が高い事例と正解ラベルとのペアを不安定な事例とする。また、上記の条件に該当しない事例と正解ラベルとのペアについては、安定な事例とする。なお、こ

の判定に関する閾値（ f ）、（ i ）、（ h ）については、例えばユーザが任意に設定してもよい。

[0056] 一例として、閾値それぞれが $f = 4$ 、 $i = 1.0$ 、 $h = 0.8$ とすると、図 5 B の出現頻度データ S_f においては、「solvent mixture」・「General」が不安定な事例となる。同様に、図 5 C のエントロピーデータ S_h および図 5 D の自己情報量データ S_i においては、「water」・「General」が不安定な事例となる。

[0057] 次いで、第 1 機械学習モデル生成部 31 は、データ 1 D の文章に関する同じ内容の事例と、同じ正解ラベル毎について判定した安定・不安定の結果をもとに、変数（ id ）に対応するデータ（文章）の安定性を示すスコアを計算し、計算結果をスコアデータ S_d に追加する（S46）。例えば、第 1 機械学習モデル生成部 31 は、不安定な事例の数または全数に対する不安定な事例の割合を指標値とし、その指標値に応じた重み付けを行うことでスコアの計算を行う。

[0058] 配列（ I ）内のデータ 1 D が空と判定された場合（S42：Yes）、第 1 機械学習モデル生成部 31 は、スコアデータ S_d をもとに、安定性の低い文章を除外した残りの文章のデータセットを第 1 機械学習モデル M1 を生成するための訓練データセット D1 とする処理を行う（S50）。

[0059] 具体的には、第 1 機械学習モデル生成部 31 は、スコアデータ S_d をソートし、スコアが低い不安定なデータ（文章）を訓練データセット D から除外する（S51）。次いで、第 1 機械学習モデル生成部 31 は、残ったデータセットを訓練データセット D1 として出力し（S52）、処理を終了する。なお、第 1 機械学習モデル生成部 31 は、スコアが低い不安定なデータ（文章）を除外する以外に、文章に含まれる一部の事例（例えば不安定な事例と判定された事例と正解ラベルとのペア）を選択して除外してもよい。

[0060] なお、前述した S30、S40 については別の処理（別の選択方法）を行うことで、訓練データセット D の中から第 1 機械学習モデル M1 を生成するための訓練データセット D1 を選択してもよい。

[0061] 具体的には、第1機械学習モデル生成部31は、収集した各事例と正解ラベルとのペアの安定性を表すスコアの初期値として、それぞれの自己情報量を設定し、以下の（・）の手順を予め指定した回数繰り返す。次いで、第1機械学習モデル生成部31は、残った訓練データセットを第1機械学習モデルM1用の訓練データセットD1とする。

- ・各文章について、出現する各事例のスコアの総和をその文章のスコアとし、スコアが最大となる文章を安定性の低い文章として除外する。

- ・除外した文章に含まれる事例のスコアを、「自己情報量／（N＋1）」とする（Nは、これまでに除外した全ての文章中で該当事例が出現した回数）

。

[0062] この別の選択方法において、第1機械学習モデル生成部31は、予め指定した回数繰り返す代わりに、各文章のスコアの中の最大値が、予め指定した閾値を下回るまで繰り返してもよい。

[0063] 上述した別の選択方法では、除外した文章に含まれる事例のスコアを下げるにより、同じ事例を含む文章が除外されにくくなる。すなわち、同じ事例が、除外される側の文章と残される側の文章の両方に含まれるようにする。なお、スコアの計算方法について、上記の例ではN＋1で自己情報量を割っているが、除外される度にスコアが小さくなるようなスコア更新方法であれば、いずれの計算方法であってもよい。

[0064] 図7は、訓練データ安定性判定処理の変形例を示すフローチャートであり、上記の別の選択方法の一例である。図7に示すように、処理が開始されると、第1機械学習モデル生成部31は、（出現）頻度を集計する処理（S20）と、自己情報量を計算する処理（S30）とを行った後、別の選択方法に関する処理（S30a）を行う。

[0065] 具体的には、第1機械学習モデル生成部31は、訓練データセットD中のデータIDの集合を処理用の配列（I）等に格納する（S41）。次いで、第1機械学習モデル生成部31は、配列（I）内のデータIDが空であるか否かを判定し（S42）、空と判定（S42：Yes）されるまでS43～

S 4 6 a の処理を繰り返す。

[0066] 配列 (I) 内のデータ I D が空でないと判定された場合 (S 4 2 : N o) 、第 1 機械学習モデル生成部 3 1 は、配列 (I) からデータ I D を 1 つ取得し、処理用の変数 (i d) に格納する (S 4 3) 。このとき、第 1 機械学習モデル生成部 3 1 は、取得したデータ I D を配列 (I) から消去する。

[0067] 次いで、第 1 機械学習モデル生成部 3 1 は、訓練データセット D 中の変数 (i d) に対応するデータから事例と正解ラベルとのペアを取得する (S 4 4) 。すなわち、第 1 機械学習モデル生成部 3 1 は、データ I D の文章に関する同じ内容の事例と、同じ正解ラベル毎のペアを取得する。次いで、第 1 機械学習モデル生成部 3 1 は、前述したスコアの計算方法を用いて各事例と正解ラベルとのペアにおけるスコア S_i を求め、その総和をスコアデータ S_d に加える (S 4 6 a) 。

[0068] 配列 (I) 内のデータ I D が空と判定された場合 (S 4 2 : Y e s) 、第 1 機械学習モデル生成部 3 1 は、スコアデータ S_d が最大のデータ d を訓練データセット D から除外する (S 5 3) 。次いで、第 1 機械学習モデル生成部 3 1 は、除外したデータ d 中の各事例と正解ラベルとのペアに対応するスコア S_i を更新し (S 5 4) 、前述した繰り返しの終了条件を満たすか否かを判定する (S 5 5) 。

[0069] 繰り返しの終了条件 (例えば、予め指定した回数繰り返す、文章のスコアの中の最大値が予め指定した閾値を下回る等) を満たさない場合 (S 5 5 : N o) 、第 1 機械学習モデル生成部 3 1 は、S 4 1 に処理を戻す。繰り返しの終了条件を満たす場合 (S 5 5 : Y e s) 、第 1 機械学習モデル生成部 3 1 は、残ったデータセットを訓練データセット D 1 として出力し (S 5 6) 、処理を終了する。

[0070] 図 4 に戻り、第 1 機械学習モデル生成部 3 1 は、S 1 0 に次いで、複数の判定手法の中からいずれかの判定手法を選択する判定手法選択処理を行う (S 1 1) 。具体的には、判定手法選択処理では、S 1 0 において前述した複数通りの選択方法のうち、どの方法を採用するかを決定する。なお、判定手

法選択処理は、S 1 0において前述した複数通りの選択方法を実施している場合に行われる処理であり、S 1 0において一通りの選択方法を実施している場合にはスキップされる。

[0071] 図8は、判定手法選択処理の概要を説明する説明図である。図8に示すように、判定手法選択処理では、訓練データセットDをk個のサブセット($D_1, \dots, D_{k-1}, D_k$)に分割し(S 7 1)、k-1個のサブセットで学習して第1機械学習モデルM1を生成する(S 7 2)。次いで、残り1個のサブセットに第1機械学習モデルM1を適用して得られた判定結果と正解とを比較し(S 7 3)、各文章のスコアを計算してソートする(正解との一致率、正解のスコア等)。次いで、ソートした結果と、複数の判定手法による判定結果とを比較し、average precision等を用いて最適な判定手法を選択する。

[0072] 図9は、判定手法選択処理の一例を示すフローチャートである。図9に示すように、処理が開始されると、第1機械学習モデル生成部31は、訓練データセットDをk個のサブセットに分割する(S 6 1)。次いで、第1機械学習モデル生成部31は、 $\{D_1 \dots D_{k-1}\}$ で第1機械学習モデルM1を生成し、生成した第1機械学習モデルM1に D_k を適用する(S 6 2)。

[0073] 次いで、第1機械学習モデル生成部31は、適用結果をもとに D_k の各データのスコアを計算してソートする(S 6 3)。次いで、第1機械学習モデル生成部31は、各訓練データにおける安定性判定手法(S 1 0における選択方法)それぞれの結果と比較し、一致の度合をスコア化する(S 6 4)。次いで、第1機械学習モデル生成部31は、S 1 0において実施した複数通りの選択方法の中で、最も一致の度合の高い手法(選択方法)の結果を採用する(S 6 5)。

[0074] 図4に戻り、S 1 1に次いで、第1機械学習モデル生成部31は、訓練データセットD1に含まれる複数の事例を用いた機械学習により第1機械学習モデルM1を生成し(S 1 2)、生成した第1機械学習モデルM1の関する第1機械学習モデル情報21を記憶部20に格納する。

[0075] 次いで、訓練データ生成部32は、第1機械学習モデル情報21をもとに第1機械学習モデルM1を構築し、訓練データセットDに含まれる各事例を構築した第1機械学習モデルM1に入力した場合に第1機械学習モデルM1が出力する判定結果を訓練データセットDに追加する（S13）。これにより、訓練データ生成部32は、訓練データセットD3を生成する。

[0076] ここで、第2機械学習モデルM2の訓練データセットD3を生成する際にノイズを加える場合を説明する。図10Aは、判定結果の追加に関する処理例を示すフローチャートであり、第1機械学習モデルM1が出力する結果にノイズを加える場合の一例である。

[0077] 図10Aに示すように、処理が開始されると、訓練データ生成部32は、訓練データセットD中のデータIDの集合を処理用の配列（I）等に格納する（S81）。次いで、訓練データ生成部32は、配列（I）内のデータIDが空であるか否かを判定し（S82）、空と判定（S82：Yes）されるまでS83～S86の処理を繰り返す。

[0078] 配列（I）内のデータIDが空でないと判定された場合（S82：No）、訓練データ生成部32は、配列（I）からデータIDを1つ取得し、処理用の変数（id）に格納する（S83）。このとき、訓練データ生成部32は、取得したデータIDを配列（I）から消去する。

[0079] 次いで、訓練データ生成部32は、訓練データセットD中の変数（id）に対応するデータに第1機械学習モデルM1を適用する（S84）。次いで、訓練データ生成部32は、第1機械学習モデルM1より得られた判定結果について、各単語（事例）に割り当てられる各ラベルのスコアをランダムに変化させる（S85）。次いで、訓練データ生成部32は、変化後のスコアをもとに、各単語に割り当てるラベルを決定する（S86）。

[0080] 図10Bは、結果データの一例を説明する説明図である。図10Bの結果データK1は、S85によりスコアをランダムに変化させた後にラベルを決定した場合のデータ例である。

[0081] 図10Bに示すように、結果データK1では、第1機械学習モデルM1よ

り得られた判定結果に含まれる推定スコアにランダムなノイズが加えられているため、スコア値が変化している。このため、S 8 6において、一部の事例では、スコアを変化させない場合とは異なる判定結果が得られる。例えば、「m i x t u r e」では、スコアの変化前では「I - G e n e r a l」と判定されるところ、スコアの変化により「O」と判定されている。

[0082] 図 1 0 Aに戻り、配列 (I) 内のデータ I Dが空と判定された場合 (S 8 2 : Y e s)、訓練データ生成部 3 2は、各事例について決定したラベルを訓練データセット Dに追加した訓練データセット D 3を出力し (S 8 4)、処理を終了する。

[0083] 図 1 1 Aは、判定結果の追加に関する処理例を示すフローチャートであり、第 1 機械学習モデル M 1 が出力する結果にノイズを加える場合の一例である。

[0084] 図 1 1 Aに示すように、処理が開始されると、訓練データ生成部 3 2は、訓練データセット D中のデータ I Dの集合を処理用の配列 (I) 等に格納する (S 8 1)。次いで、訓練データ生成部 3 2は、配列 (I) 内のデータ I Dが空であるか否かを判定し (S 8 2)、空と判定 (S 8 2 : Y e s) されるまで S 8 3 ~ S 8 6 aの処理を繰り返す。

[0085] 配列 (I) 内のデータ I Dが空でないと判定された場合 (S 8 2 : N o)、訓練データ生成部 3 2は、配列 (I) からデータ I Dを 1 つ取得し、処理用の変数 (i d) に格納する (S 8 3)。このとき、訓練データ生成部 3 2は、取得したデータ I Dを配列 (I) から消去する。

[0086] 次いで、訓練データ生成部 3 2は、訓練データセット D中の変数 (i d) に対応するデータに第 1 機械学習モデル M 1 を適用する (S 8 4)。次いで、訓練データ生成部 3 2は、第 1 機械学習モデル M 1 より得られた判定結果について、各単語 (事例) に割り当てられる各ラベルのスコアを確率値に変換する (S 8 5 a)。具体的には、スコアが高ければ選択されやすくなるような確率値となるように、スコアに応じた確率値に変換する。次いで、訓練データ生成部 3 2は、変換した確率値をもとに、各単語に割り当てるラベル

を決定する（S 8 6 a）。

[0087] 図 1 1 B は、結果データの一例を説明する説明図である。図 1 1 B の結果データ K 2 は、スコアから変換した後の確率値をもとにラベルを決定した場合のデータ例である。

[0088] 図 1 0 B に示すように、結果データ K 2 では、確率値に変換された推定スコアをもとに、確率的にラベルを決定（選択）している。このため、確率値が均衡する一部の事例では、スコアの大小を元にした判定結果とは異なる判定結果が得られる場合がある。例えば、「p r o p y l」では、スコアの大小では「l - M o l e c u l a r」と判定されるところ、確率的な選択により「B - M o l e c u l a r」とされている。

[0089] 図 1 2 A は、判定結果の追加に関する処理例を示すフローチャートであり、第 1 機械学習モデル M 1 の入力にノイズを加える場合の一例である。

[0090] 図 1 2 A に示すように、処理が開始されると、訓練データ生成部 3 2 は、訓練データセット D 中のデータ I D の集合を処理用の配列（l）等に格納する（S 8 1）。次いで、訓練データ生成部 3 2 は、配列（l）内のデータ I D が空であるか否かを判定し（S 8 2）、空と判定（S 8 2 : Y e s）されるまで S 8 3 ~ S 8 4 c の処理を繰り返す。

[0091] 配列（l）内のデータ I D が空でないと判定された場合（S 8 2 : N o）、訓練データ生成部 3 2 は、配列（l）からデータ I D を 1 つ取得し、処理用の変数（i d）に格納する（S 8 3）。このとき、訓練データ生成部 3 2 は、取得したデータ I D を配列（l）から消去する。

[0092] 次いで、訓練データ生成部 3 2 は、訓練データセット D 中の変数（i d）に対応するデータの一部の単語をランダムに選択し、選択した単語を別の単語に置換する（S 8 4 a）。なお、置換対象の単語は、データの中からランダムに選んだり、推定結果の確信度（スコア）をもとに選んだりしてもよい。また、別の単語への置換は、任意の単語への置換であってもよい。または、類義語・関連語辞書を利用して置換対象の単語の類義語・関連語への置換であってもよいし、単語分散表現を利用して選択した単語への置換であって

もよい。

[0093] 次いで、訓練データ生成部32は、置換後のデータに第1機械学習モデルM1を適用し（S84b）、第1機械学習モデルM1より得られた判定結果より各単語に割り当てるラベルを決定する（S84c）。

[0094] 図12Bは、結果データの一例を説明する説明図である。図12Bの結果データK3は、置換後の単語（2列目の単語）をもとにラベルを決定した場合のデータ例である。

[0095] 図12Bに示すように、結果データK3では、一部の事例（単語）において、その内容が別の内容に置換されている。例えば、上から6段目の「m i x t u r e」は「c o m p o u n d」に置換されている。このように、第1機械学習モデルM1に入力するデータに対してノイズを付加する構成であってもよい。

[0096] 以上のように、情報処理装置1は、第1機械学習モデル生成部31と、訓練データ生成部32とに関する処理を実行する制御部30を有する。第1機械学習モデル生成部31は、訓練データセットDに含まれる各事例の出現頻度に基づいて、訓練データセットDから複数の事例を選択する。また、第1機械学習モデル生成部31は、選択した複数の事例を用いた機械学習により第1機械学習モデルM1を生成する。訓練データ生成部32は、訓練データセットDと、訓練データセットDに含まれる各事例を入力した場合に第1機械学習モデルM1が出力する結果とを合わせた訓練データセットD3を生成する。また、制御部30は、訓練データセットD3を用いて第2機械学習モデルM2を生成する第2機械学習モデル生成部33に関する処理を実行する。制御部30では、分類対象のデータを分類する分類タスクにおいて、第1機械学習モデルM1に分類対象のデータを入力し、第1機械学習モデルM1の出力結果を得る。次いで、制御部30では、第1機械学習モデルM1の出力結果を第2機械学習モデルM2に入力して、第2機械学習モデルM2より分類結果を得る。これにより、単一の機械学習モデルの分類精度よりも精度の良い分類結果を得ることができる。

[0097] このように、情報処理装置 1 では、訓練データセット D に含まれる各事例の出現頻度に基づいて選択した複数の事例を用いた機械学習により第 1 機械学習モデル M 1 を生成するので、第 2 機械学習モデル M 2 を学習するための訓練データセット D 3 を生成する際に、例えば第 1 機械学習モデル M 1 を k 回繰り返して生成することがない。したがって、情報処理装置 1 では、第 2 機械学習モデル M 2 を学習するための訓練データセット D 3 を効率的に生成することができ、効率的な機械学習を実行することができる。

[0098] また、第 1 機械学習モデル生成部 3 1 は、訓練データセット D のうち出現頻度が閾値未満の事例を選択対象から除外する。このように、情報処理装置 1 では、訓練データセット D うち、出現頻度が閾値未満であり、第 1 機械学習モデル M 1 による判定結果が不安定になると推定される事例を選択対象から除外した上で第 1 機械学習モデル M 1 を生成する。このため、訓練データセット D に含まれる各事例を入力した場合に第 1 機械学習モデル M 1 が出力する結果において、不安定になると推定される事例では、訓練データセット D の正解ラベルとは異なる結果が得られやすくなる。したがって、情報処理装置 1 では、第 1 機械学習モデル M 1 の判定結果の誤りを修正するように第 2 機械学習モデル M 2 を生成するための訓練データセット D 3 を生成することができ、第 2 機械学習モデル M 2 による最終判定結果の精度を向上させることができる。

[0099] また、第 1 機械学習モデル生成部 3 1 は、出現頻度に基づいて各事例のエントロピーおよび自己情報量を計算し、訓練データセット D のうち、自己情報量が閾値より大きく、かつ、エントロピーが閾値未満の事例を選択対象から除外する。このように、情報処理装置 1 では、訓練データセット D のうち、自己情報量が閾値より大きく、かつ、エントロピーが閾値未満の事例であり、第 1 機械学習モデル M 1 による判定結果が不安定になると推定される事例を選択対象から除外した上で第 1 機械学習モデル M 1 を生成する。このため、訓練データセット D に含まれる各事例を入力した場合に第 1 機械学習モデル M 1 が出力する結果において、不安定になると推定される事例では、訓

練データセットDの正解ラベルとは異なる結果が得られやすくなる。したがって、情報処理装置1では、第1機械学習モデルM1の判定結果の誤りを修正するように第2機械学習モデルM2を生成するための訓練データセットD3を生成することができ、第2機械学習モデルM2による最終判定結果の精度を向上させることができる。

[0100] また、訓練データ生成部32は、訓練データセットDと、訓練データセットDに含まれる各事例の一部の事例の内容を変更した後のデータセットに含まれる各事例を入力した場合に第1機械学習モデルM1が出力する結果とを合わせて第2機械学習モデルM2用の訓練データセットD3を生成する。このように、訓練データセットDに含まれる各事例の一部の事例の内容を変更し、訓練データセットDにノイズを加えることで、第1機械学習モデルM1の判定結果が変化しやすい事例では、第1機械学習モデルM1が出力する結果において、訓練データセットDの正解ラベルとは異なる結果が得られやすくなる。したがって、情報処理装置1では、第1機械学習モデルM1の判定結果の誤りを修正するように第2機械学習モデルM2を生成するための訓練データセットD3を生成することができ、第2機械学習モデルM2による最終判定結果の精度を向上させることができる。

[0101] また、訓練データ生成部32は、第1機械学習モデルM1が出力する結果に特定の割合でノイズを加えて訓練データセットD3を生成する。このように、情報処理装置1では、第1機械学習モデルM1が出力する結果に特定の割合でノイズを加え、第1機械学習モデルM1の判定結果の誤りを修正するように第2機械学習モデルM2を生成するための訓練データセットD3を生成してもよい。

[0102] また、制御部30は、第2機械学習モデル生成部33に関する処理を実行する。第2機械学習モデル生成部33は、生成した訓練データセットD3に基づいた機械学習により第2機械学習モデルM2を生成する。これにより、情報処理装置1では、生成した訓練データセットD3により第2機械学習モデルM2を生成することができる。

[0103] また、訓練データセットDに含まれる各事例は、教師とする複数の文章それぞれに含まれる単語である。これにより、情報処理装置1では、文章に含まれる各単語の品詞推定、固有表現抽出、語義判定等を最終結果として出力する第2機械学習モデルM2を生成するための訓練データセットD3を効率的に生成することができる。

[0104] なお、図示した各装置の各構成要素は、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。

[0105] また、情報処理装置1で行われる各種処理機能は、CPU（またはMPU、MCU（Micro Controller Unit）等のマイクロ・コンピュータ）あるいはGPU（Graphics Processing Unit）上で、その全部または任意の一部を実行するようにしてもよい。また、各種処理機能は、CPU（またはMPU、MCU等のマイクロ・コンピュータ）あるいはGPUで解析実行されるプログラム上、またはワイヤードロジックによるハードウェア上で、その全部または任意の一部を実行するようにしてもよいことは言うまでもない。また、情報処理装置1で行われる各種処理機能は、クラウドコンピューティングにより、複数のコンピュータが協働して実行してもよい。

[0106] ところで、上記の実施形態で説明した各種の処理は、予め用意されたプログラムをコンピュータで実行することで実現できる。そこで、以下では、上記の実施形態と同様の機能を有するプログラムを実行するコンピュータ（ハードウェア）の一例を説明する。図13は、コンピュータ構成の一例を示すブロック図である。

[0107] 図13に示すように、コンピュータ200は、各種演算処理を実行するCPU201と、画像処理や機械学習処理等の所定の演算処理に特化したGPU201aと、データ入力を受け付ける入力装置202と、モニタ203と、スピーカ204とを有する。また、コンピュータ200は、記憶媒体から

プログラム等を読み取る媒体読取装置 205 と、各種装置と接続するためのインタフェース装置 206 と、有線または無線により外部機器と通信接続するための通信装置 207 とを有する。また、コンピュータ 200 は、各種情報を一時記憶する RAM 208 と、ハードディスク装置 209 とを有する。また、コンピュータ 200 内の各部（201～209）は、バス 210 に接続される。

[0108] ハードディスク装置 209 には、上記の実施形態で説明した制御部 30 における第 1 機械学習モデル生成部 31、訓練データ生成部 32 および第 2 機械学習モデル生成部 33 等における各種の処理を実行するためのプログラム 211 が記憶される。また、ハードディスク装置 209 には、プログラム 211 が参照する訓練データセット D 等の各種データ 212 が記憶される。入力装置 202 は、例えば、操作者から操作情報の入力を受け付ける。モニタ 203 は、例えば、操作者が操作する各種画面を表示する。インタフェース装置 206 は、例えば印刷装置等が接続される。通信装置 207 は、LAN（Local Area Network）等の通信ネットワークと接続され、通信ネットワークを介した外部機器との間で各種情報をやりとりする。

[0109] CPU 201 あるいは GPU 201a は、ハードディスク装置 209 に記憶されたプログラム 211 を読み出して、RAM 208 に展開して実行することで、第 1 機械学習モデル生成部 31、訓練データ生成部 32 および第 2 機械学習モデル生成部 33 等に関する各種の処理を行う。なお、プログラム 211 は、ハードディスク装置 209 に記憶されていなくてもよい。例えば、コンピュータ 200 が読み取り可能な記憶媒体に記憶されたプログラム 211 を読み出して実行するようにしてもよい。コンピュータ 200 が読み取り可能な記憶媒体は、例えば、CD-ROM や DVD ディスク、USB（Universal Serial Bus）メモリ等の可搬型記録媒体、フラッシュメモリ等の半導体メモリ、ハードディスクドライブ等が対応する。また、公衆回線、インターネット、LAN 等に接続された装置にこのプログラム 211 を記憶させておき、コンピュータ 200 がこれらからプログラム 211 を読み出して実

行するようにしてもよい。

符号の説明

[0110] 1 …情報処理装置

1 0 …入出力部

2 0 …記憶部

2 1 …第 1 機械学習モデル情報

2 2 …第 2 機械学習モデル情報

3 0 …制御部

3 1 …第 1 機械学習モデル生成部

3 2 …訓練データ生成部

3 3 …第 2 機械学習モデル生成部

2 0 0 …コンピュータ

2 0 1 …CPU

2 0 1 a …GPU

2 0 2 …入力装置

2 0 3 …モニタ

2 0 4 …スピーカ

2 0 5 …媒体読取装置

2 0 6 …インタフェース装置

2 0 7 …通信装置

2 0 8 …RAM

2 0 9 …ハードディスク装置

2 1 0 …バス

2 1 1 …プログラム

2 1 2 …各種データ

D、D 1 ～D 3、D 1 0 0、D 1 0 1 …訓練データセット

K 1 ～K 3 …結果データ

M 1、M 1 0 1 …第 1 機械学習モデル

M₂、M₁₀₂…第2機械学習モデル

S_f…出現頻度データ

S_h…エントロピーデータ

S_i…自己情報量データ

S_d…スコアデータ

請求の範囲

- [請求項1] 第1の訓練データ群に含まれる同一のラベルが付された同一のデータの出現頻度に基づいて、前記第1の訓練データ群から複数のデータを選択し、
- 選択した前記複数のデータを用いた機械学習により第1の機械学習モデルを生成し、
- 前記第1の訓練データ群と、前記第1の訓練データ群に含まれるデータを入力した場合に前記第1の機械学習モデルが出力する結果とを合わせた第2の訓練データ群を生成する、
- 処理をコンピュータに実行させることを特徴とする機械学習プログラム。
- [請求項2] 前記選択する処理は、前記第1の訓練データ群のうち前記出現頻度が閾値未満のデータを選択対象から除外する処理を含む、
- ことを特徴とする請求項1に記載の機械学習プログラム。
- [請求項3] 前記選択する処理は、前記出現頻度に基づいて前記同一のラベルが付された同一のデータのエントロピーおよび自己情報量を計算し、前記第1の訓練データ群のうち、前記自己情報量が第1の閾値より大きく、かつ、前記エントロピーが第2の閾値未満のデータを選択対象から除外する処理を含む、
- ことを特徴とする請求項1に記載の機械学習プログラム。
- [請求項4] 前記第2の訓練データ群を生成する処理は、前記第1の訓練データ群と、前記第1の訓練データ群に含まれる第1のデータの内容を変更して生成された第2のデータを入力した場合に前記第1の機械学習モデルが出力する第1の結果とを合わせて前記第2の訓練データ群を生成する処理を含む、
- ことを特徴とする請求項1に記載の機械学習プログラム。
- [請求項5] 前記第2の訓練データ群を生成する処理は、前記第1の訓練データ群と、前記第1の訓練データ群に含まれる第1のデータを入力した場

合に前記第 1 の機械学習モデルが出力する第 1 の結果の内容を変更して生成された第 2 の結果とを合わせて前記第 2 の訓練データ群を生成する処理を含む、

ことを特徴とする請求項 1 に記載の機械学習プログラム。

[請求項6] 生成した前記第 2 の訓練データ群に基づいた機械学習により第 2 の機械学習モデルを生成する、
処理を前記コンピュータに実行させることを特徴とする請求項 1 に記載の機械学習プログラム。

[請求項7] 第 1 の訓練データ群に含まれる同一のラベルが付された同一のデータの出現頻度に基づいて、前記第 1 の訓練データ群から複数のデータを選択し、

選択した前記複数のデータを用いた機械学習により第 1 の機械学習モデルを生成し、

前記第 1 の訓練データ群と、前記第 1 の訓練データ群に含まれるデータを入力した場合に前記第 1 の機械学習モデルが出力する結果とを合わせた第 2 の訓練データ群を生成する、

処理をコンピュータが実行することを特徴とする機械学習方法。

[請求項8] 前記選択する処理は、前記第 1 の訓練データ群のうち前記出現頻度が閾値未満のデータを選択対象から除外する処理を含む、

ことを特徴とする請求項 7 に記載の機械学習方法。

[請求項9] 前記選択する処理は、前記出現頻度に基づいて前記同一のラベルが付された同一のデータのエントロピーおよび自己情報量を計算し、前記第 1 の訓練データ群のうち、前記自己情報量が第 1 の閾値より大きく、かつ、前記エントロピーが第 2 の閾値未満のデータを選択対象から除外する処理を含む、

ことを特徴とする請求項 7 に記載の機械学習方法。

[請求項10] 前記第 2 の訓練データ群を生成する処理は、前記第 1 の訓練データ群と、前記第 1 の訓練データ群に含まれる第 1 のデータの内容を変更

して生成された第2のデータを入力した場合に前記第1の機械学習モデルが出力する第1の結果とを合わせて前記第2の訓練データ群を生成する処理を含む、

ことを特徴とする請求項7に記載の機械学習方法。

[請求項11] 前記第2の訓練データ群を生成する処理は、前記第1の訓練データ群と、前記第1の訓練データ群に含まれる第1のデータを入力した場合に前記第1の機械学習モデルが出力する第1の結果の内容を変更して生成された第2の結果とを合わせて前記第2の訓練データ群を生成する処理を含む、

ことを特徴とする請求項7に記載の機械学習方法。

[請求項12] 生成した前記第2の訓練データ群に基づいた機械学習により第2の機械学習モデルを生成する、

ことを特徴とする請求項7に記載の機械学習方法。

[請求項13] 第1の訓練データ群に含まれる同一のラベルが付された同一のデータの出現頻度に基づいて、前記第1の訓練データ群から複数のデータを選択し、

選択した前記複数のデータを用いた機械学習により第1の機械学習モデルを生成し、

前記第1の訓練データ群と、前記第1の訓練データ群に含まれるデータを入力した場合に前記第1の機械学習モデルが出力する結果とを合わせた第2の訓練データ群を生成する、

処理を実行する制御部を含むことを特徴とする情報処理装置。

[請求項14] 前記選択する処理は、前記第1の訓練データ群のうち前記出現頻度が閾値未満のデータを選択対象から除外する処理を含む、

ことを特徴とする請求項13に記載の情報処理装置。

[請求項15] 前記選択する処理は、前記出現頻度に基づいて前記同一のラベルが付された同一のデータのエントロピーおよび自己情報量を計算し、前記第1の訓練データ群のうち、前記自己情報量が第1の閾値より大き

く、かつ、前記エントロピーが第2の閾値未満のデータを選択対象から除外する処理を含む、

ことを特徴とする請求項13に記載の情報処理装置。

[請求項16] 前記第2の訓練データ群を生成する処理は、前記第1の訓練データ群と、前記第1の訓練データ群に含まれる第1のデータの内容を変更して生成された第2のデータを入力した場合に前記第1の機械学習モデルが出力する第1の結果とを合わせて前記第2の訓練データ群を生成する処理を含む、

ことを特徴とする請求項13に記載の情報処理装置。

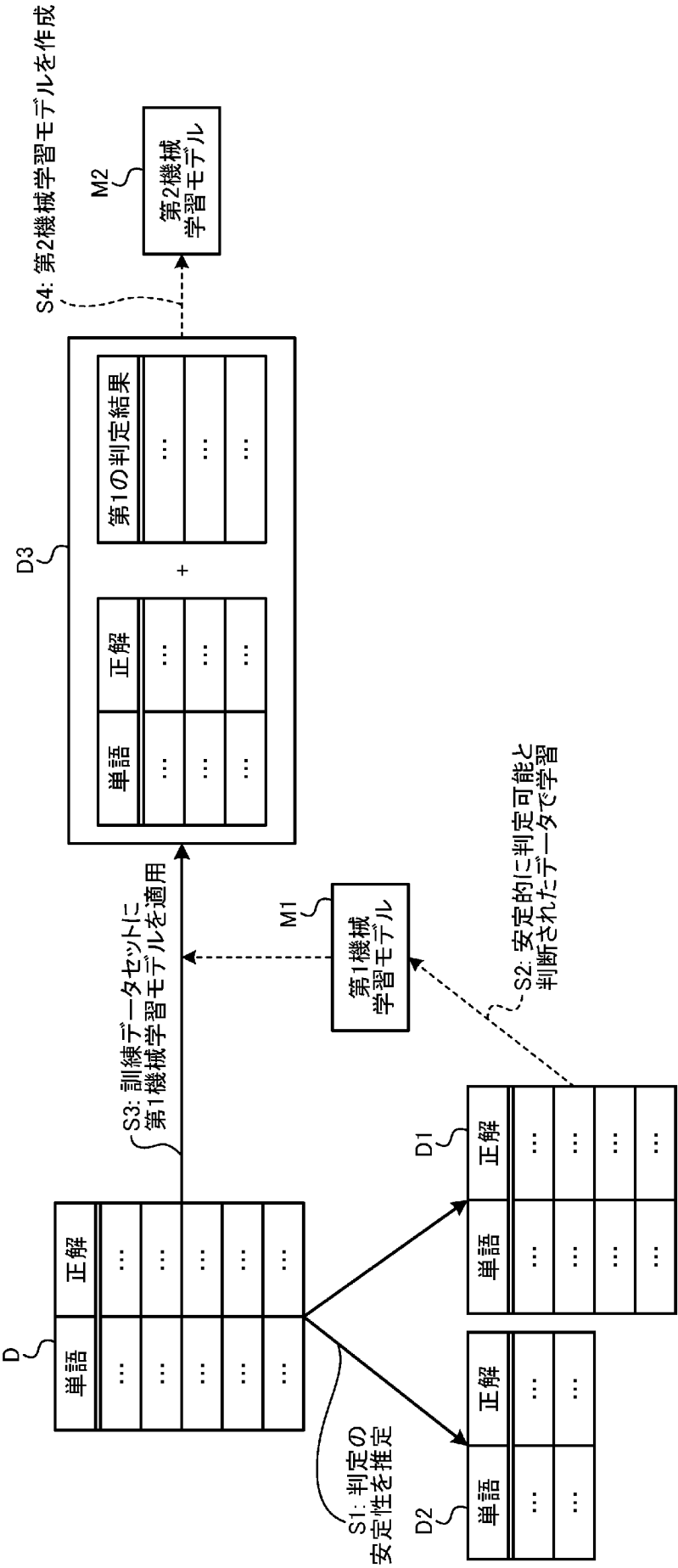
[請求項17] 前記第2の訓練データ群を生成する処理は、前記第1の訓練データ群と、前記第1の訓練データ群に含まれる第1のデータを入力した場合に前記第1の機械学習モデルが出力する第1の結果の内容を変更して生成された第2の結果とを合わせて前記第2の訓練データ群を生成する処理を含む、

ことを特徴とする請求項13に記載の情報処理装置。

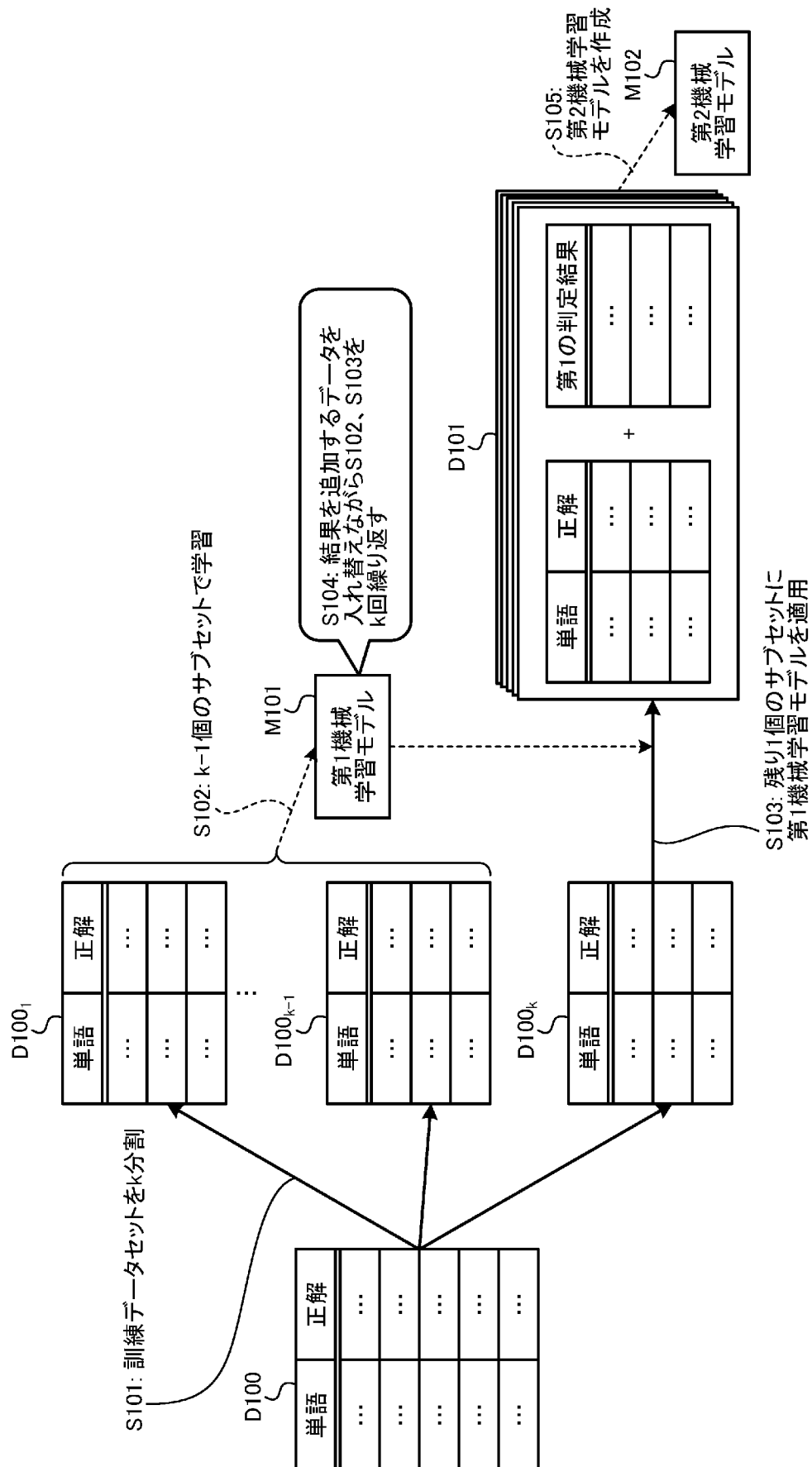
[請求項18] 生成した前記第2の訓練データ群に基づいた機械学習により第2の機械学習モデルを生成する処理をさらに前記制御部が実行する、

ことを特徴とする請求項13に記載の情報処理装置。

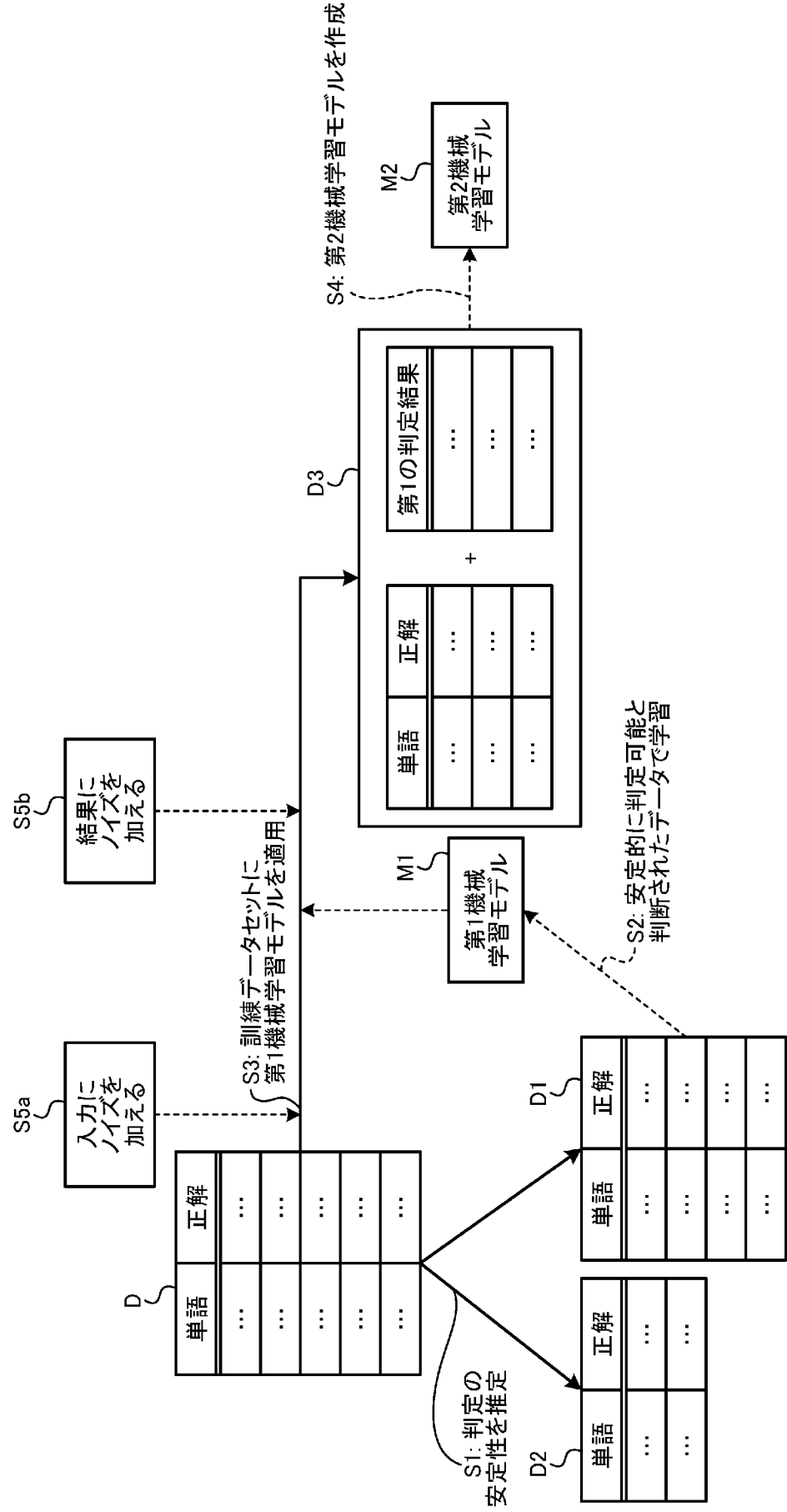
[図1]



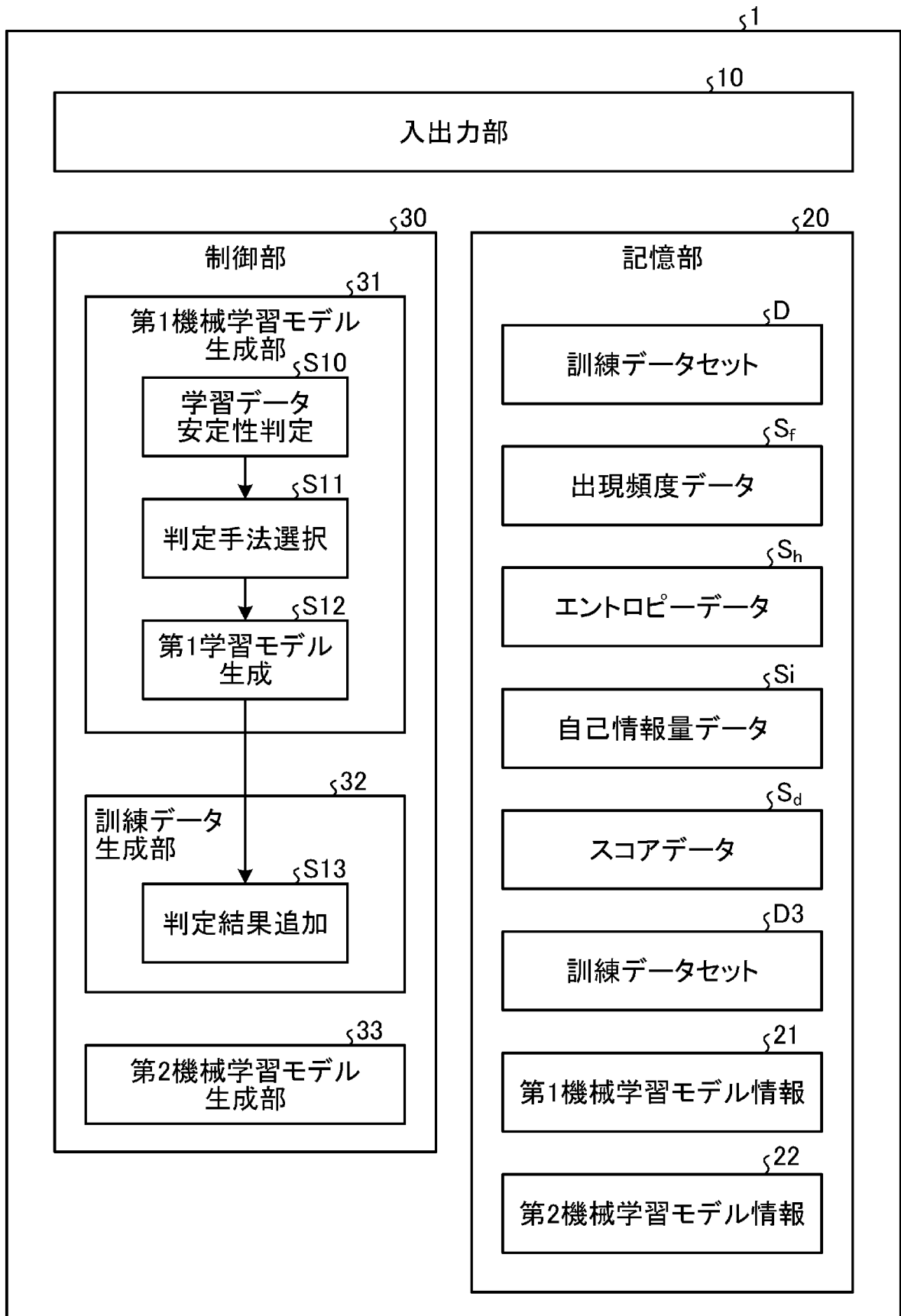
[図2]



[図3]



[図4]



[図5A]

D
└─┘

データID	単語	正解ラベル
1	The	O
1	invention	O
1	provides	O
1	a	O
1	solvent	B-General
1	mixture	I-General
1	comprising	O
⋮	⋮	⋮
2	An	O
2	improved	O
2	catalyst	B-General
⋮	⋮	⋮

[図5B]

S_f

事例	正解・頻度ペア集合
solvent mixture	{{(General, 3)}}
n-propyl bromide	{{(Molecular, 5)}}
water	{{(Molecular, 2083), (General, 5)}}
⋮	⋮

[図5C]

S_h

事例	エントロピー
solvent mixture	1.233
n-propyl bromide	0.650
water	0.619
⋮	⋮

[図5D]

S_i

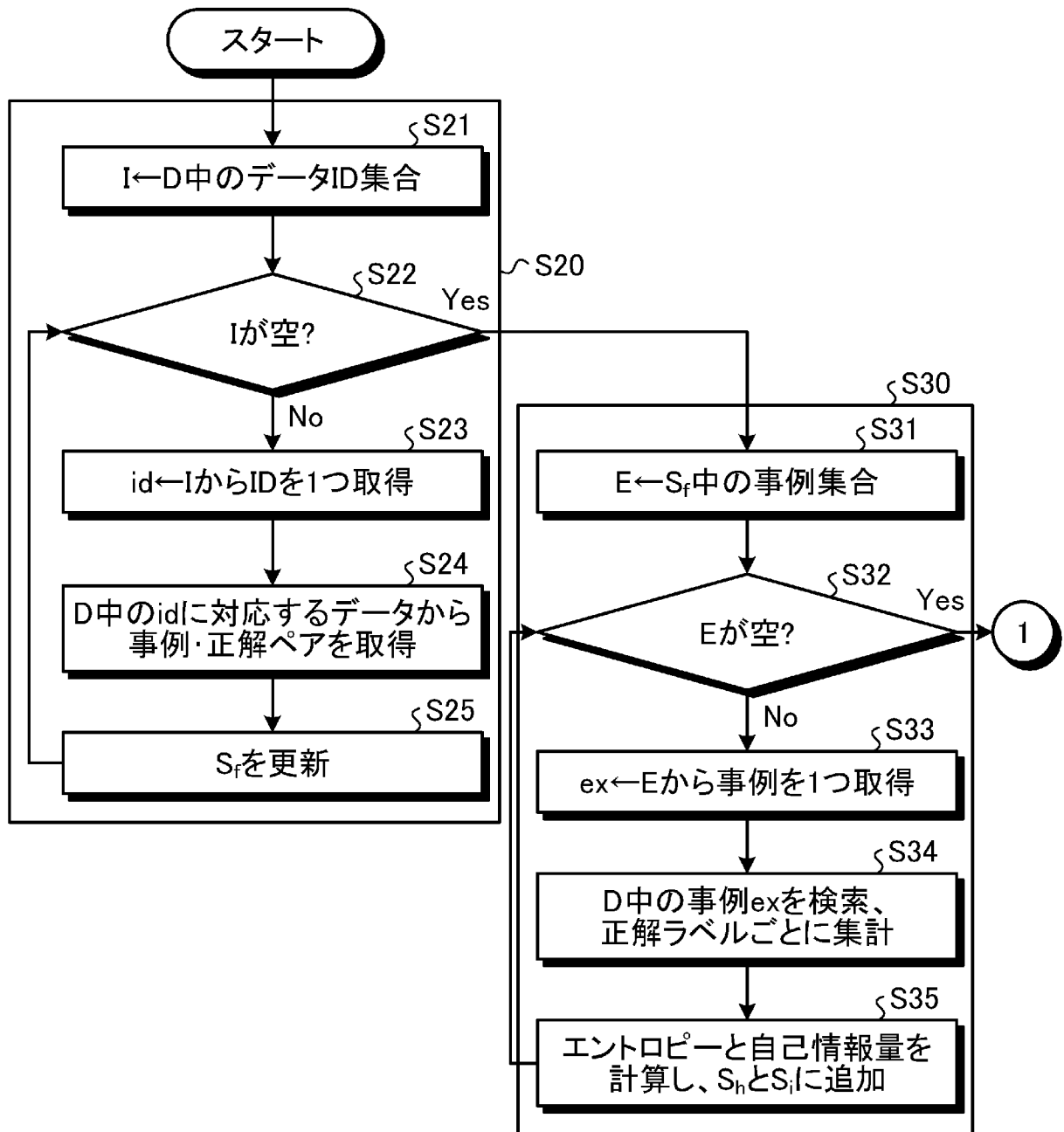
事例	正解	自己情報量
solvent mixture	General	2.585
n-propyl bromide	Molecular	0.263
water	General	8.930
water	Molecular	0.228
⋮	⋮	⋮

[図5E]

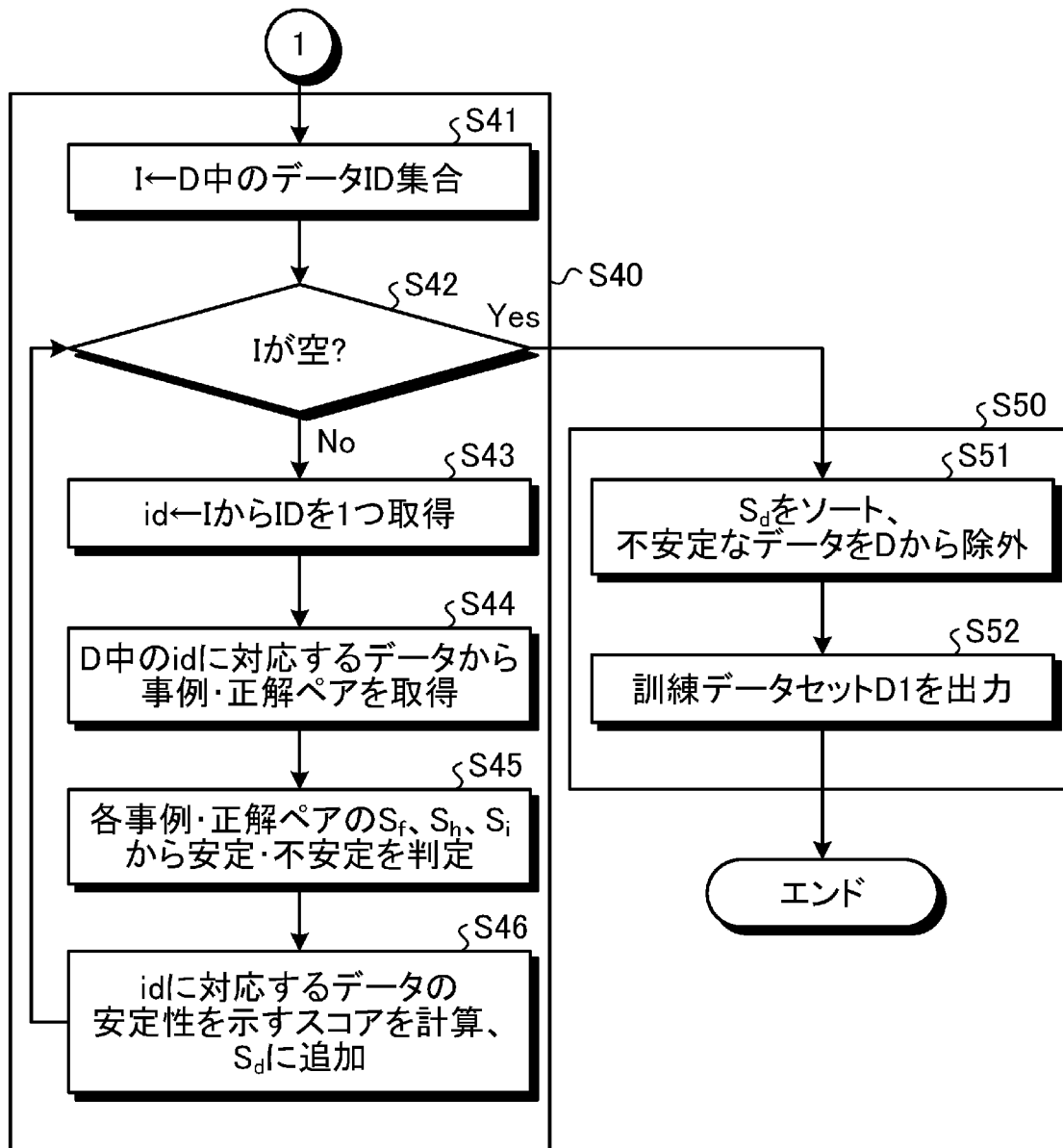
S_d

データID	スコア
1	3.0
2	1.0
⋮	⋮

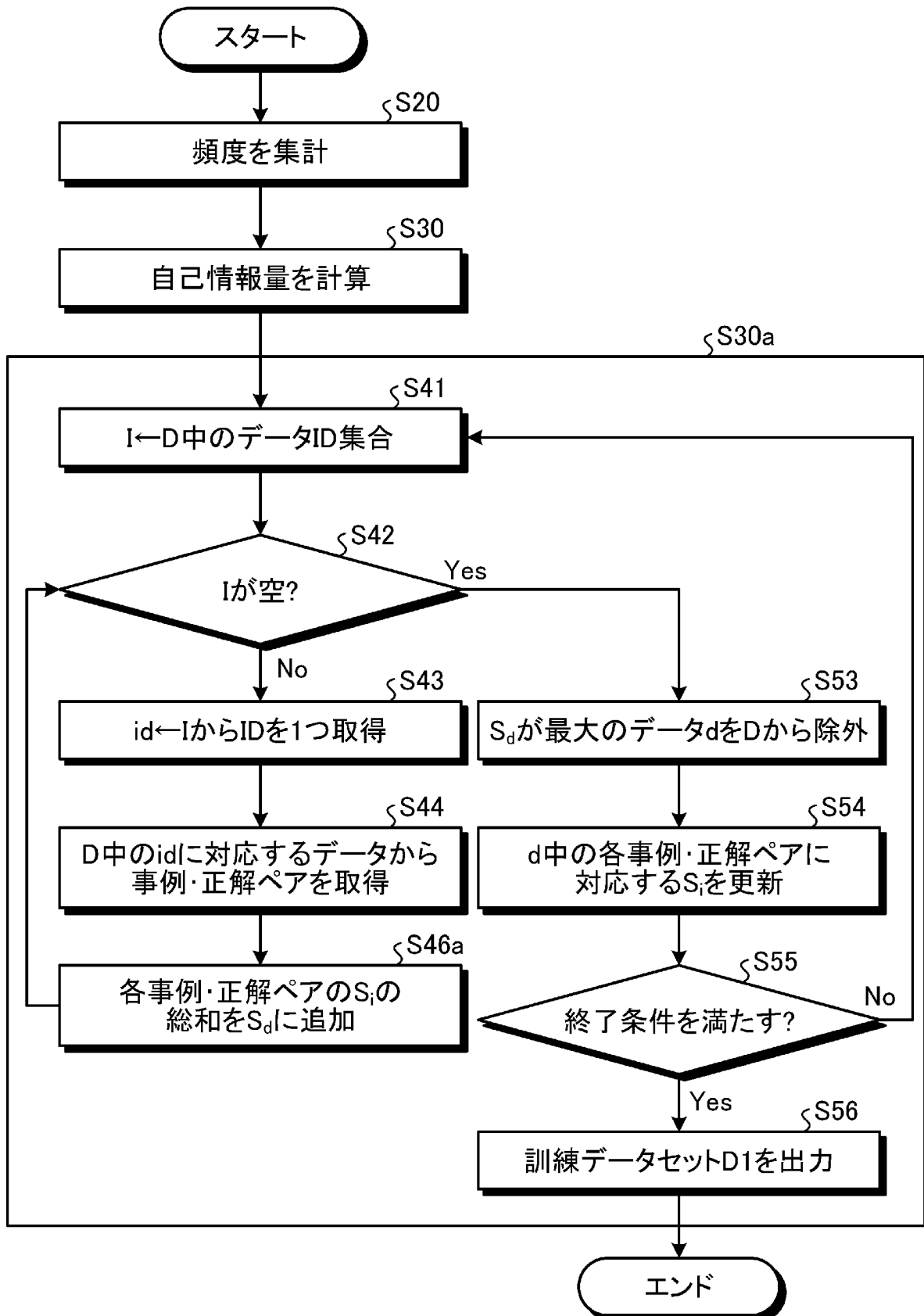
[図6A]



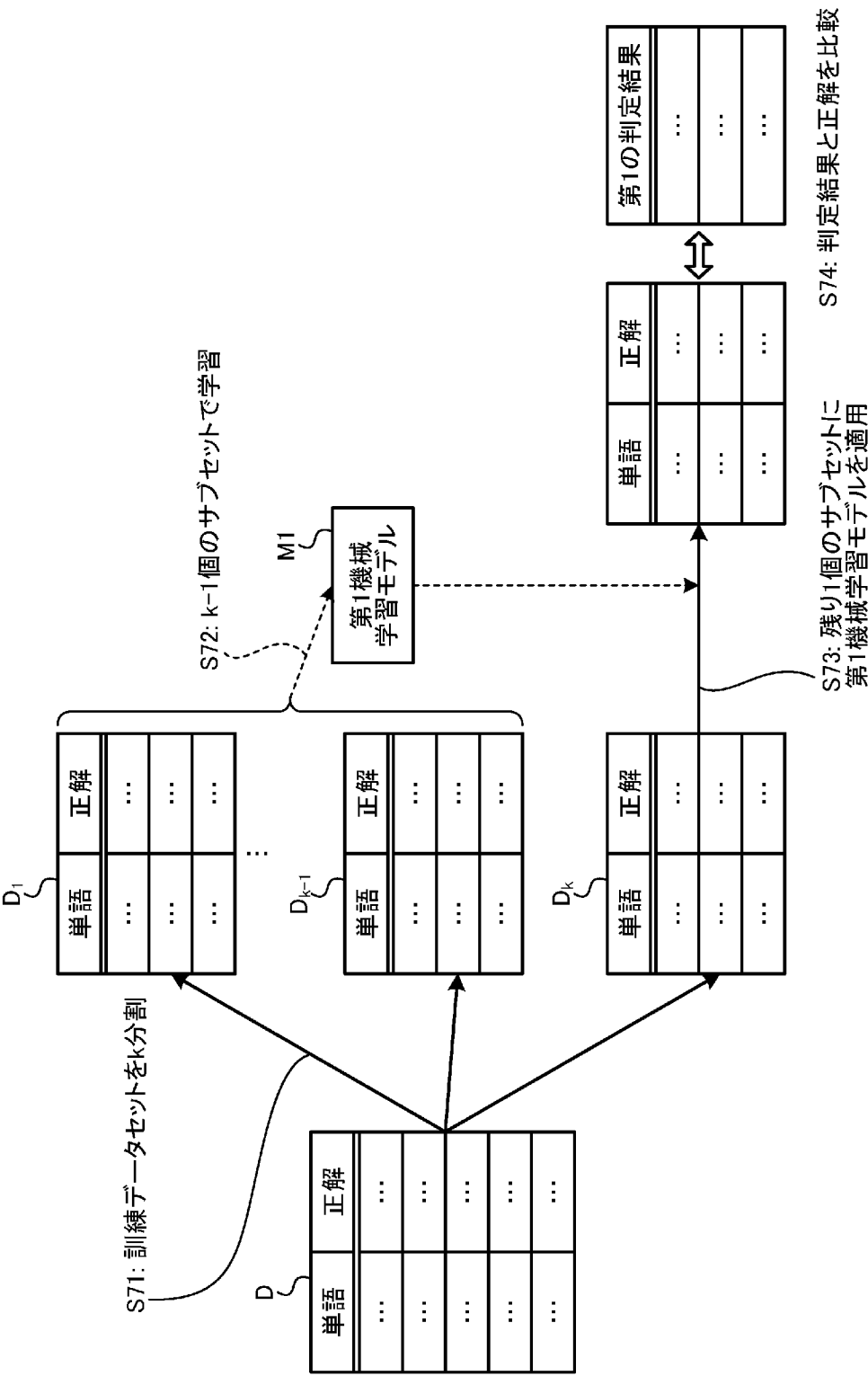
[図6B]



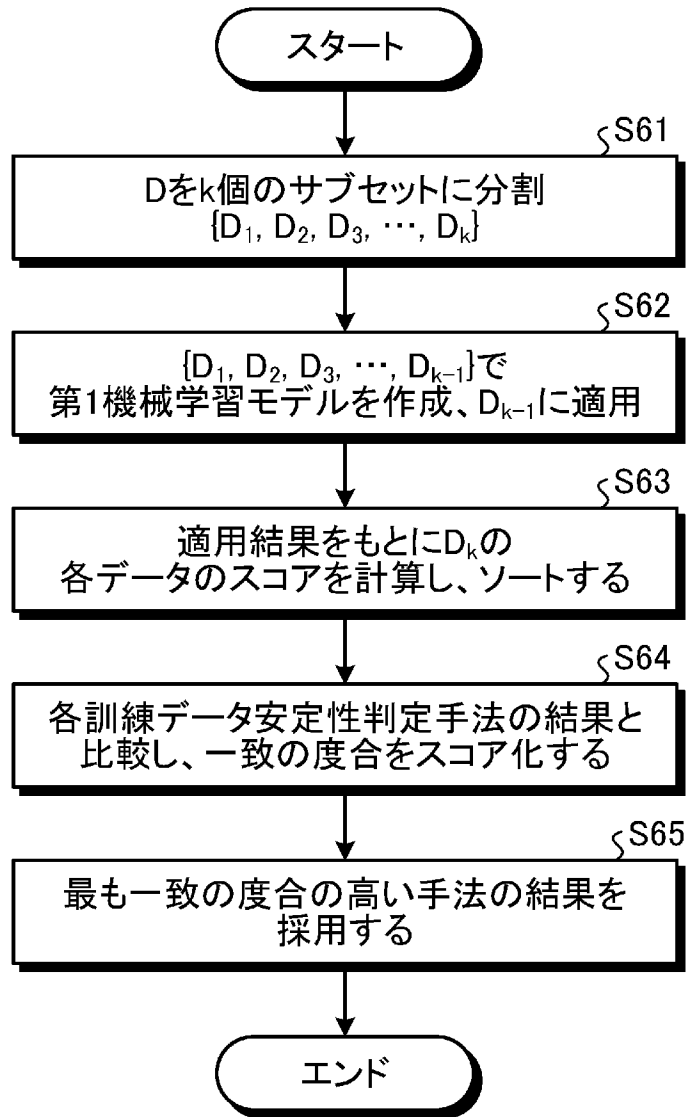
[図7]



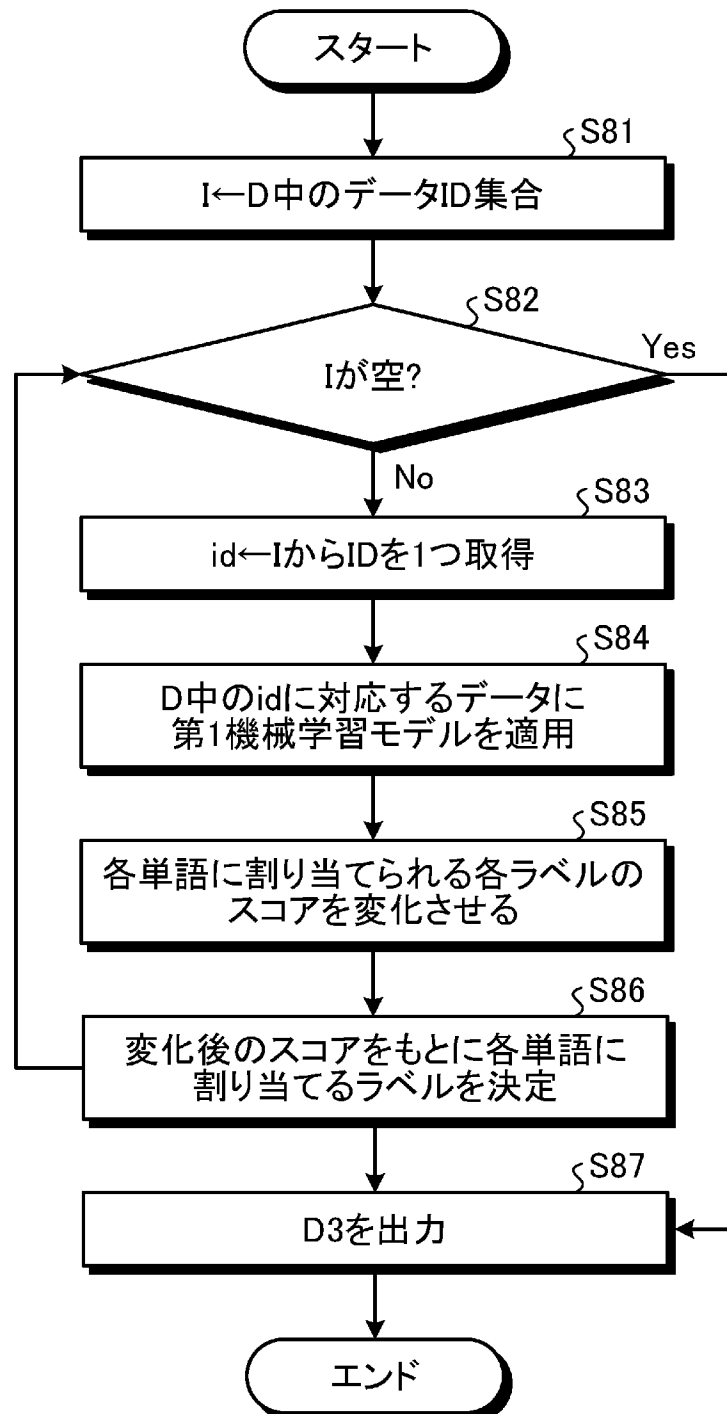
[図8]



[図9]



[図10A]

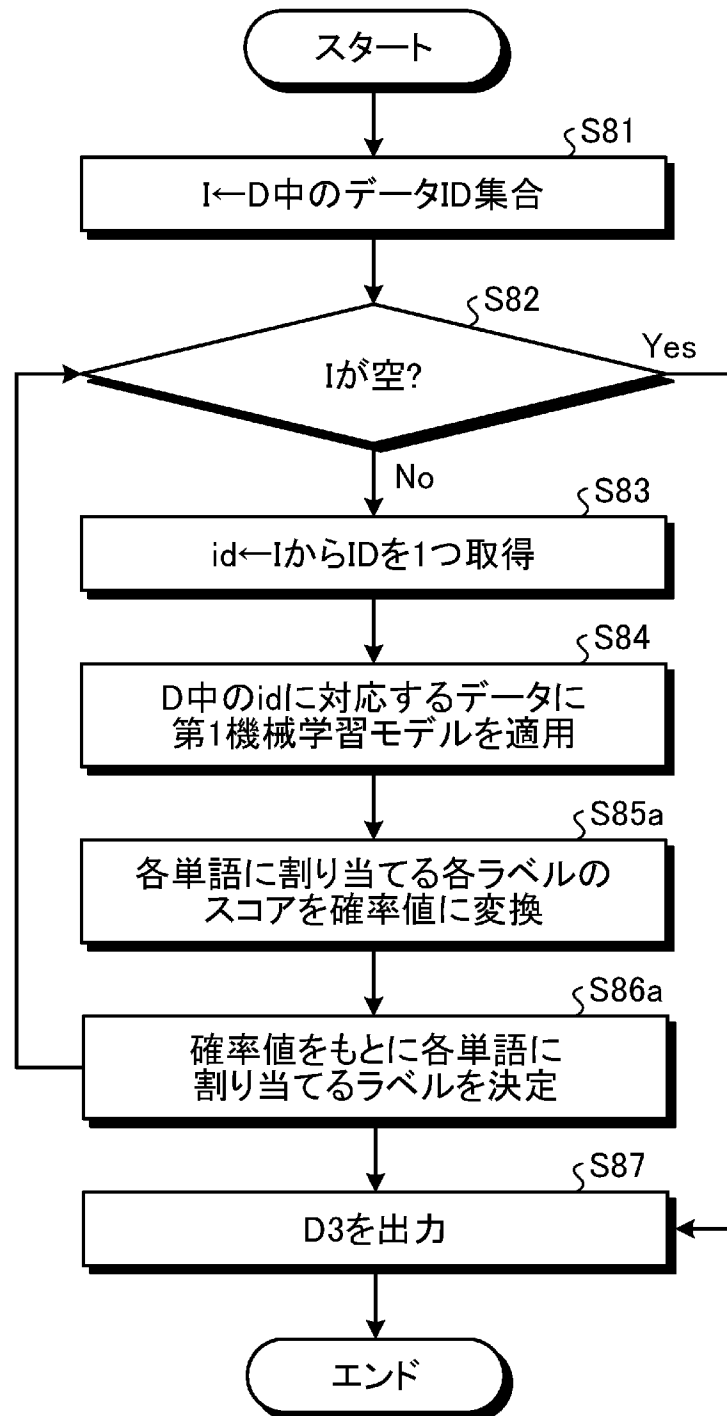


[図10B]

K1

単語	推定スコア	ノイズ追加	判定
The	O: 1.0	O: 1.0	O
invention	O: 1.0	O: 1.0	O
provides	O: 1.0	O: 1.0	O
a	O: 0.9, B-General: 0.1	O: 0.89, B-General: 0.11	O
solvent	B-General: 0.9, I-General: 0.1	B-General: 0.85, I-General: 0.15	B-General
mixture	I-General: 0.35, O: 0.34, B-General: 0.31	I-General: 0.33, O: 0.35, B-General: 0.32	O
comprising	O: 1.0	O: 1.0	O
n	B-Molecular: 0.9, O:0.1	B-Molecular: 0.88, O:0.12	B-Molecular
—	I-Molecular: 0.8, O:0.2	I-Molecular: 0.77, O:0.23	I-Molecular
propyl	I-Molecular: 0.6, B-Molecular: 0.4	I-Molecular: 0.55, B-Molecular: 0.45	I-Molecular
bromide	I-Molecular: 0.7, B-Molecular: 0.3	I-Molecular: 0.69, B-Molecular: 0.21	I-Molecular
,	O: 1.0	O: 1.0	O
:	:	:	:

[図11A]

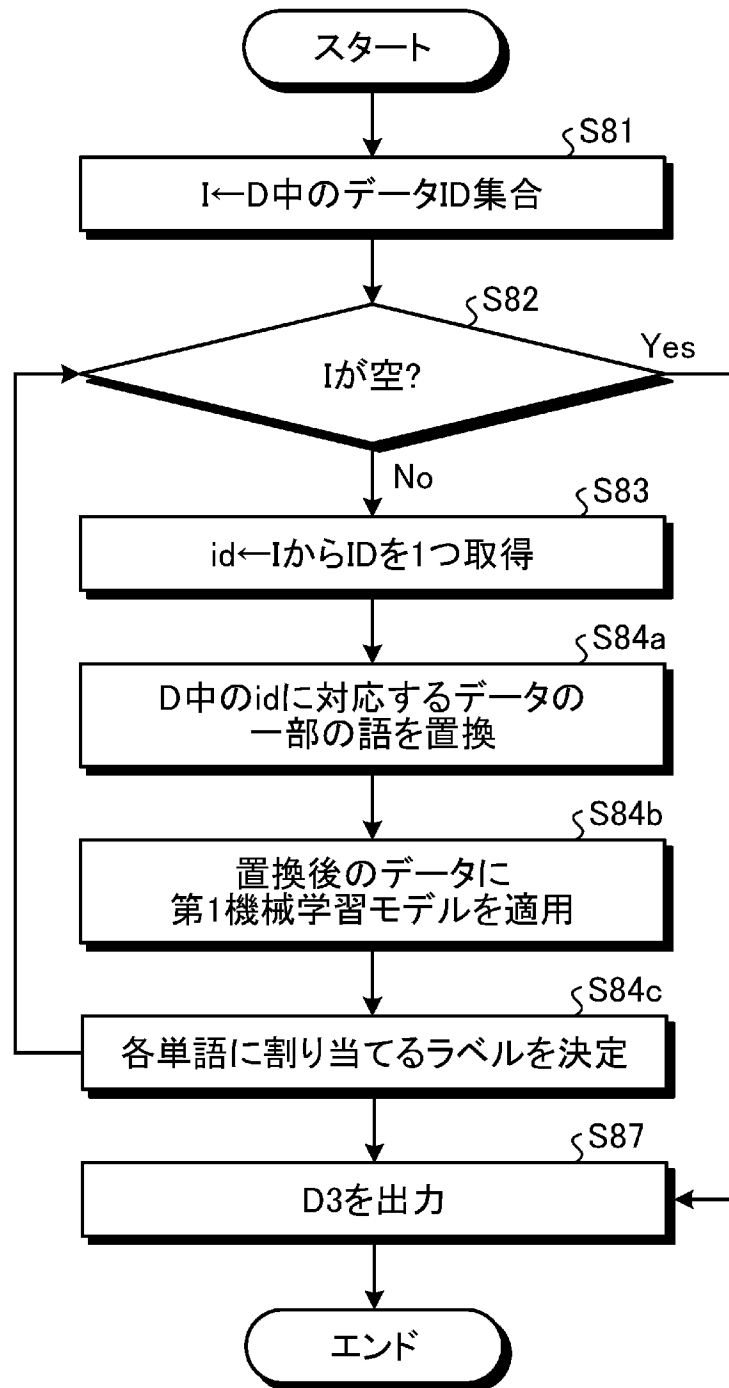


[図11B]

K2

単語	推定スコア	確率的に選択
The	O: 1.0	O
invention	O: 1.0	O
provides	O: 1.0	O
a	O: 0.9, B-General: 0.1	O
solvent	B-General: 0.9, I-General: 0.1	B-General
mixture	I-General: 0.35, O: 0.34, B-General: 0.31	I-General
comprising	O: 1.0	O: 1.0
n	B-Molecular: 0.9, O:0.1	B-Molecular
–	I-Molecular: 0.8, O:0.2	I-Molecular
propyl	I-Molecular: 0.6, B-Molecular: 0.4	B-Molecular
bromide	I-Molecular: 0.7, B-Molecular: 0.3	I-Molecular
,	O: 1.0	O: 1.0
⋮	⋮	⋮

[図12A]

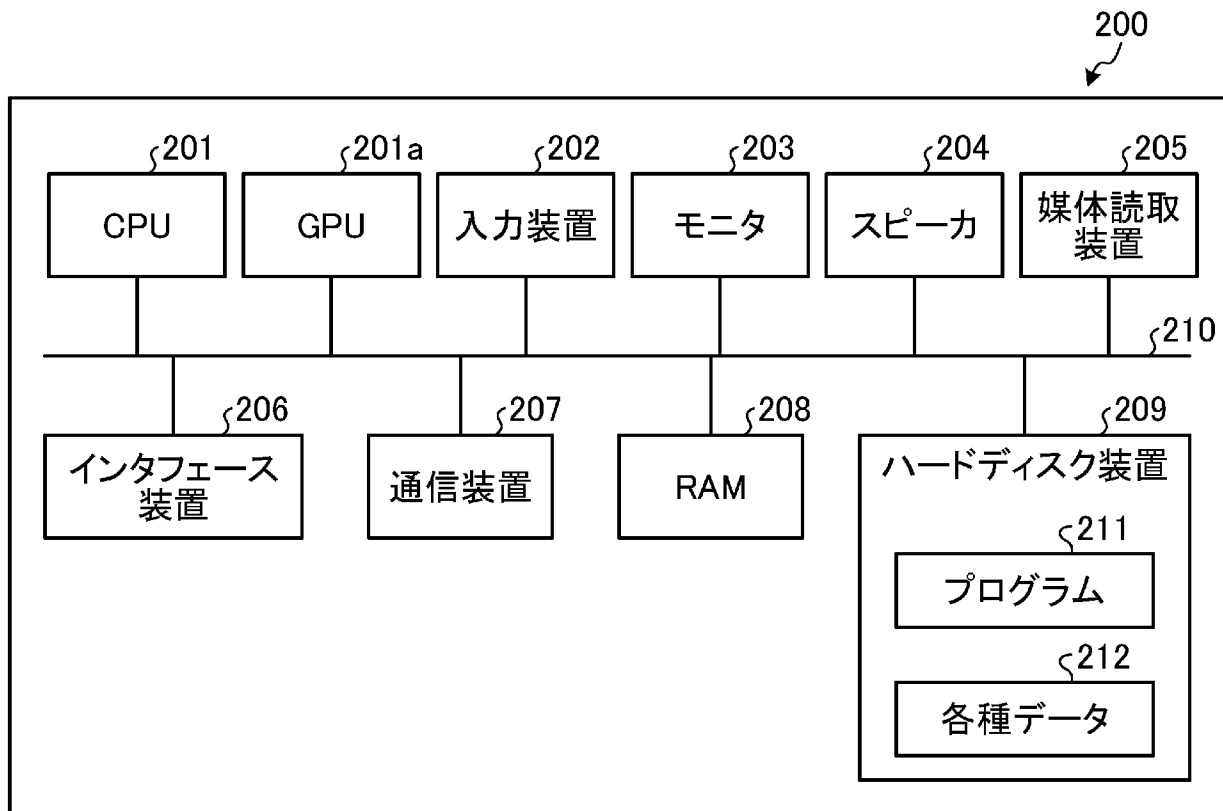


[図12B]

K3

単語		推定スコア	判定
The	The	O: 1.0	O
invention	invention	O: 1.0	O
provides	provides	O: 1.0	O
a	a	O: 0.9, B-General: 0.1	O
solvent	solvent	B-General: 0.9, I-General: 0.1	B-General
mixture	compound	B-General: 0.5, I-General: 0.3, O: 0.2	B-General
comprising	comprising	O: 1.0	O: 1.0
n	n	B-Molecular: 0.9, O:0.1	B-Molecular
–	–	I-Molecular: 0.8, O:0.2	I-Molecular
propyl	ethyl	I-Molecular: 0.7, B-Molecular: 0.3	I-Molecular
bromide	bromide	I-Molecular: 0.7, B-Molecular: 0.3	I-Molecular
,	,	O: 1.0	O: 1.0
⋮	⋮	⋮	⋮

[図13]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/027411

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G06N20/00 (2019.01) i

FI: G06N20/00 130

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)

Int. Cl. G06N20/00

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 1971-2020

Registered utility model specifications of Japan 1996-2020

Published registered utility model applications of Japan 1994-2020

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 2019-016025 A (HITACHI, LTD.) 31 January 2019, abstract, fig. 3	1-18
A	JP 2006-252333 A (NARA INSTITUTE OF SCIENCE AND TECHNOLOGY et al.) 21 September 2006, abstract, fig. 5	1-18
A	JP 2018-045559 A (FUJITSU LTD.) 22 March 2018, abstract, fig. 1	1-18
A	JP 2020-047079 A (FUJITSU LTD.) 26 March 2020, abstract, fig. 16	1-18



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

29.09.2020

Date of mailing of the international search report

06.10.2020

Name and mailing address of the ISA/

Japan Patent Office

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/JP2020/027411

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
JP 2019-016025 A	31.01.2019	US 2019/0012611 A1 abstract, fig. 3	
JP 2006-252333 A	21.09.2006	WO 2006/095853 A1 abstract, fig. 5 CN 101138001 A	
JP 2018-045559 A	22.03.2018	US 2018/0082215 A1 abstract, fig. 1	
JP 2020-047079 A	26.03.2020	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 20/00(2019.01)i FI: G06N20/00 130		
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）） G06N20/00		
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 - 2020年 日本国実用新案登録公報 1996 - 2020年 日本国登録実用新案公報 1994 - 2020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 2019-016025 A（株式会社日立製作所）31.01.2019（2019 - 01 - 31） 要約及び図3	1-18
A	JP 2006-252333 A（国立大学法人奈良先端科学技術大学院大学 外2名）21.09.2006 （2006 - 09 - 21） 要約及び図5	1-18
A	JP 2018-045559 A（富士通株式会社）22.03.2018（2018 - 03 - 22） 要約及び図1	1-18
A	JP 2020-047079 A（富士通株式会社）26.03.2020（2020 - 03 - 26） 要約及び図16	1-18
☐ 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、一般的な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 29.09.2020		国際調査報告の発送日 06.10.2020
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） 坂庭 剛史 5B 9288 電話番号 03-3581-1101 内線 3545

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/JP2020/027411

引用文献			公表日	パテントファミリー文献	公表日
JP	2019-016025	A	31.01.2019	US 2019/0012611 A1 要約及び図3	
JP	2006-252333	A	21.09.2006	WO 2006/095853 A1 要約及び図5 CN 101138001 A	
JP	2018-045559	A	22.03.2018	US 2018/0082215 A1 要約及び図1	
JP	2020-047079	A	26.03.2020	(ファミリーなし)	