

(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12月22日(22.12.2022)



(10) 国際公開番号

WO 2022/264232 A1

- (51) 国際特許分類:
G06F 40/44 (2020.01) *G06F 40/45* (2020.01)
- (21) 国際出願番号: PCT/JP2021/022581
- (22) 国際出願日: 2021年6月14日(14.06.2021)
- (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: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).
- (72) 発明者: 鈴木 敏 (SUZUKI, Satoshi); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP). 永田

昌明 (NAGATA, Masaaki); 〒1808585 東京都武蔵野市緑町3丁目9-11 N T T 知的財産センタ内 Tokyo (JP).

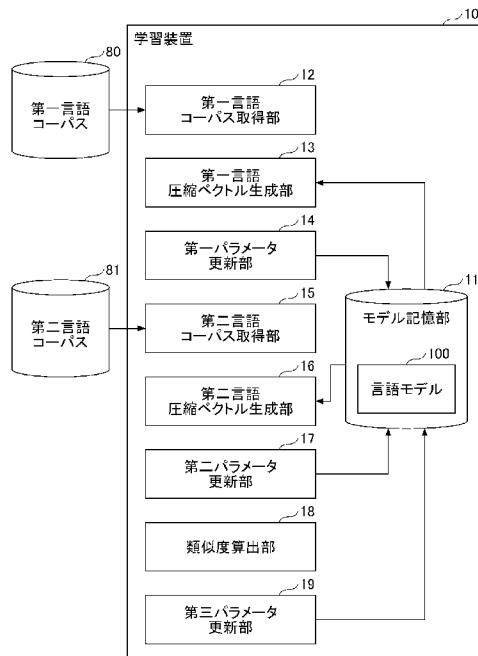
(74) 代理人: 伊東 忠重, 外 (ITO, Tadashige et al.); 〒1000005 東京都千代田区丸の内二丁目1番1号 丸の内 M Y P L A Z A (明治安田生命ビル) 16階 Tokyo (JP).

(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,

(54) Title: LEARNING DEVICE, PARALLEL TRANSLATION SENTENCE OUTPUT DEVICE, LEARNING METHOD, PARALLEL TRANSLATION SENTENCE OUTPUT METHOD, AND PROGRAM

(54) 発明の名称: 学習装置、対訳文出力装置、学習方法、対訳文出力方法およびプログラム

[図1]



- 10 Learning device
11 Model storage unit
12 First-language corpus acquisition unit
13 First-language compressed vector generation unit
14 First parameter update unit
15 Second-language corpus acquisition unit
16 Second-language compressed vector generation unit
17 Second parameter update unit
18 Similarity degree calculation unit
19 Third parameter update unit
80 First-language corpus
81 Second-language corpus
100 Language model

(57) Abstract: A learning device for updating parameters of a language model, said learning device being provided with: a first-language compressed vector generation unit which generates a first compressed vector by compressing a first-language corpus on the basis of the language model; a first parameter update unit which updates the parameters of the language model on the basis of the first compressed vector; a second-language compressed vector generation unit which



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, 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))

generates a second compressed vector by compressing a second-language corpus on the basis of the language model; a second parameter update unit which updates the parameters of the language model on the basis of the second compressed vector; a similarity degree calculation unit which calculates the degree of similarity between a first compressed vector and a second compressed vector generated respectively from texts having the same meaning; and a third parameter update unit which updates the parameters of the language model on the basis of the degree of similarity.

(57) 要約：言語モデルのパラメータを更新するための学習装置であって、前記言語モデルに基づいて、第一言語のコーパスを圧縮した第一圧縮ベクトルを生成する第一言語圧縮ベクトル生成部と、前記第一圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第一パラメータ更新部と、前記言語モデルに基づいて、第二言語のコーパスを圧縮した第二圧縮ベクトルを生成する第二言語圧縮ベクトル生成部と、前記第二圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第二パラメータ更新部と、それぞれ意味が同じテキストから生成された前記第一圧縮ベクトルと前記第二圧縮ベクトルの類似度を算出する類似度算出部と、前記類似度に基づいて、言語モデルのパラメータを更新する第三パラメータ更新部と、を備える学習装置である。

明 細 書

発明の名称：

学習装置、対訳文出力装置、学習方法、対訳文出力方法およびプログラム
技術分野

[0001] 本発明は、学習装置、対訳文出力装置、学習方法、対訳文出力方法およびプログラムに関する。

背景技術

[0002] 自然言語処理による自動翻訳を、対訳データを用いて実現する技術が知られている。例えば、日英翻訳であれば、日本語文書と英語文書との対応を大量に学習させることで翻訳精度を上げる方法が知られている。

[0003] 非特許文献1には、意味が類似する文を類似度の高いベクトルへと写像する技術が開示されている。

先行技術文献

非特許文献

[0004] 非特許文献1：Nils Reimers, SentenceTransformers Documentation, [online], インターネット<URL: <https://www.sbert.net/index.html>>

発明の概要

発明が解決しようとする課題

[0005] 上記の従来技術において利用できる対訳データは、類似したテキストの集合、或いは、対訳テキストのペアの集合であるが、テキスト同士の対応付けが明確でないものも多く、それらに関しては利用を諦めるか、あるいは人手で対応付けを行うなどが必要である。そのため、利用しやすい対訳データを大量に準備することが難しいという問題がある。

[0006] 開示の技術は、対訳データによらずに、異なる言語間で意味の近いテキストを出力することを目的とする。

課題を解決するための手段

[0007] 開示の技術は、言語モデルのパラメータを更新するための学習装置であって、前記言語モデルに基づいて、第一言語のコーパスを圧縮した第一圧縮ベクトルを生成する第一言語圧縮ベクトル生成部と、前記第一圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第一パラメータ更新部と、前記言語モデルに基づいて、第二言語のコーパスを圧縮した第二圧縮ベクトルを生成する第二言語圧縮ベクトル生成部と、前記第二圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第二パラメータ更新部と、それぞれ意味が同じテキストから生成された前記第一圧縮ベクトルと前記第二圧縮ベクトルの類似度を算出する類似度算出部と、前記類似度に基づいて、言語モデルのパラメータを更新する第三パラメータ更新部と、を備える学習装置である。

発明の効果

[0008] 対訳データによらずに、異なる言語間で意味の近いテキストを出力することができる。

図面の簡単な説明

- [0009] [図1]学習装置の機能構成図である。
- [図2]言語モデルの一例を示す第一の図である。
- [図3]言語モデルの一例を示す第二の図である。
- [図4]言語モデルの一例を示す第三の図である。
- [図5]学習処理の流れの一例を示すフローチャートである。
- [図6]対訳文出力装置の機能構成図である。
- [図7]対訳文出力処理における言語モデルの使用方法について説明するための図である。
- [図8]対訳文出力処理の流れの一例を示すフローチャートである。
- [図9]コンピュータのハードウェア構成例を示す図である。

発明を実施するための形態

[0010] 以下、図面を参照して本発明の実施の形態（本実施の形態）を説明する。

以下で説明する実施の形態は一例に過ぎず、本発明が適用される実施の形態は、以下の実施の形態に限られるわけではない。

[0011] （本実施の形態の概要）

本実施の形態に係る学習装置は、二つの言語のそれぞれのテキストを一つの意味空間上にベクトルとして圧縮変換する学習と、二つの言語の対訳コーパスを用いて同じ意味の文を入力とした場合に、各言語の圧縮ベクトルを一致或いは近傍に集める学習と、を行って、言語モデルのパラメータを更新する。また、本実施の形態に係る対訳文出力装置は、学習装置によって更新されたパラメータを適用した言語モデルに基づいて対訳文を出力する。なお、対訳文とは、入力された第一言語のテキストに対応する第二言語のテキストであって、単語、フレーズ、文章、段落等を含む。

[0012] （学習装置の機能構成例）

図1は、学習装置の機能構成図である。学習装置10は、モデル記憶部11と、第一言語コーパス取得部12と、第一言語圧縮ベクトル生成部13と、第一パラメータ更新部14と、第二言語コーパス取得部15と、第二言語圧縮ベクトル生成部16と、第二パラメータ更新部17と、類似度算出部18と、第三パラメータ更新部19と、を備える。

[0013] モデル記憶部11は、言語モデル100を記憶する。言語モデル100は、第一言語（例えば日本語）と第二言語（例えば英語）のそれぞれのテキストを圧縮または復元するオートエンコーダであって、中間層に入力されたテキストに対応する圧縮ベクトルを生成するように学習される。

[0014] 第一言語コーパス取得部12は、第一言語コーパス80を取得する。第一言語コーパス80は、例えば日本語を第一言語とする場合、日本語のテキストを構造化して集積された単言語のコーパスである。

[0015] 第一言語圧縮ベクトル生成部13は、言語モデル100に基づいて第一言語のコーパスを圧縮して、圧縮ベクトル（第一圧縮ベクトル）を生成する。

[0016] 第一パラメータ更新部14は、第一言語コーパス80を使用して、生成さ

れた圧縮ベクトルがコーパスに対応する意味となるように学習することによって、言語モデル100のモデルパラメータを更新する。

[0017] 第二言語コーパス取得部15は、第二言語コーパス81を取得する。第二言語コーパス81は、例えば英語を第二言語とする場合、英語のテキストを構造化して集積された単言語のコーパスである。

[0018] 第二言語圧縮ベクトル生成部16は、言語モデル100に基づいて第二言語のコーパスを圧縮して、圧縮ベクトル（第二圧縮ベクトル）を生成する。

[0019] 第二パラメータ更新部17は、第二言語コーパス81を使用して、生成された圧縮ベクトルがコーパスに対応する意味となるように学習することによって、言語モデル100のモデルパラメータを更新する。

[0020] 類似度算出部18は、第一言語コーパス80から生成された圧縮ベクトル（第一圧縮ベクトル）と、第二言語コーパス81から生成された圧縮ベクトル（第二圧縮ベクトル）が、同じ意味のテキストから生成された場合に、これらの圧縮ベクトルの類似度を算出する。類似度は、例えば圧縮ベクトル間のノルムまたは $\cos$ 距離等である。なお、類似度算出部18は、同じ意味のテキストから生成されたか否かを、対訳コーパス等に基づいて判定する。また、第一言語圧縮ベクトル生成部13および第二言語圧縮ベクトル生成部16は、類似度の算出の対象となる第一言語の圧縮ベクトル（第一圧縮ベクトル）と第二言語の圧縮ベクトル（第二圧縮ベクトル）とを、第一言語コーパス80または第二言語コーパス81ではなく対訳コーパスを圧縮して生成しても良い。

[0021] 第三パラメータ更新部19は、同じ意味のテキストから生成される圧縮ベクトルの類似度を最大化するように、言語モデル100のモデルパラメータを更新する。

[0022] （言語モデルの構成例）

図2は、言語モデルの構成例を示す第一の図である。図2に示す言語モデル100は、（1）第一言語のオートエンコーダに相当するエンコーダ101、中間層102およびデコーダ103と、（2）第二言語のオートエンコ

ーダに相当するエンコーダ104、中間層105およびデコーダ106と、を含む。また、(3)第三パラメータ更新部19による処理の対象は、第一言語の圧縮ベクトルの生成に関わるエンコーダ101および中間層102と、第二言語の圧縮ベクトルの生成に関わるエンコーダ104および中間層105である。

[0023] エンコーダ101、104およびデコーダ103、106は、それぞれTransformers/Bert等(参考文献[1])の既存の言語モデルを適用しても良い。

[0024] 図2に示す言語モデルの形態は、圧縮ベクトル同士の類似度としてcosine距離を適用する場合などに使用でき、2つの言語の意味空間が完全な一致ではなく、緩やかに対応する場合にも用いることができる形態である。

[0025] 図3は、言語モデルの構成例を示す第二の図である。図3に示す言語モデル100は、(1)第一言語のオートエンコーダに相当するエンコーダ101、中間層102およびデコーダ103と、(2)第二言語のオートエンコーダに相当するエンコーダ104、中間層102およびデコーダ106と、を含む。すなわち、言語モデル100が備える中間層は、第一言語と第二言語で共通の中間層102となっている。また、(3)第三パラメータ更新部19による処理の対象は、第一言語の圧縮ベクトルの生成に関わるエンコーダ101および中間層102と、第二言語の圧縮ベクトルの生成に関わるエンコーダ104および中間層102である。

[0026] 図3に示す言語モデルの形態は、図2に示す言語モデルの形態とほぼ同等であるが、2つの言語の意味空間が完全に一致する場合に用いることができる。

[0027] 図4は、言語モデルの構成例を示す第三の図である。図4に示す言語モデル100は、(1)第一言語のオートエンコーダに相当するエンコーダ101、中間層102およびデコーダ103と、(2)第二言語のオートエンコーダに相当するエンコーダ101、中間層102およびデコーダ106と、を含む。すなわち、言語モデル100が備えるエンコーダおよび中間層は、

第一言語と第二言語で共通のエンコーダ 101 および中間層 102 となっている。また、(3) 第三パラメータ更新部 19 による処理の対象は、第一言語および第二言語の圧縮ベクトルの生成に関わるエンコーダ 101 および中間層 102 である。

[0028] 図 4 に示す言語モデルの形態は、例えば、図 2 に示す言語モデルの (3) に Sentence-Bert (非特許文献 1) の多言語モデルを適用した場合のモデルであり、適用するモデル内に二つのエンコーダの機能を既に含んでいる場合に適用できる形式的な形態である。図 4 に示す言語モデルの形態は、実質的には図 2 または図 3 に示す言語モデルの形態と同等である。

[0029] 同様に、二つのデコーダの機能を一つのデコーダに包含する場合には、形式的に一つのデコーダにより表現することもできる。

[0030] (学習装置の動作例)

次に、学習装置 10 の動作例について、図面を参照して説明する。学習装置 10 は、ユーザの操作等を受けて、または定期的に、学習処理を開始する。

[0031] 図 5 は、学習処理の流れの一例を示すフローチャートである。第一言語コーパス取得部 12 は、第一言語のコーパスを取得する (ステップ S 11)。次に、第一言語圧縮ベクトル生成部 13 は、第一言語のコーパスの圧縮ベクトルを生成する (ステップ S 12)。続いて、第一パラメータ更新部 14 は、第一言語のコーパスの圧縮ベクトルに基づいて、言語モデルのパラメータを更新する (ステップ S 13)。

[0032] ステップ S 11 からステップ S 13 までの処理は、第一言語のオートエンコーダを学習するための第一の学習に相当する。

[0033] また、第二言語コーパス取得部 15 は、第二言語のコーパスを取得する (ステップ S 14)。次に、第二言語圧縮ベクトル生成部 16 は、第二言語のコーパスの圧縮ベクトルを生成する (ステップ S 15)。続いて、第二パラメータ更新部 17 は、第二言語のコーパスの圧縮ベクトルに基づいて、言語モデルのパラメータを更新する (ステップ S 16)。

- [0034] ステップS 1 4 からステップS 1 6 までの処理は、第二言語のオートエンコーダを学習するための第二の学習に相当する。
- [0035] 続いて、類似度算出部 1 8 は、第一言語のコーパスの圧縮ベクトルと、第二言語のコーパスの圧縮ベクトルとの類似度を算出する（ステップS 1 7）。算出される類似度は、例えば圧縮ベクトル間のノルムまたは $\cosine$ 距離等である。
- [0036] そして、第三パラメータ更新部 1 9 は、類似度に基づいて、言語モデルのパラメータを更新する（ステップS 1 8）。例えば、第三パラメータ更新部 1 9 は、意味が共通する圧縮ベクトル間のノルムまたは $\cosine$ 距離を最小化するように、言語モデルのパラメータを更新する。
- [0037] ステップS 1 7 からステップS 1 8 までの処理は、類似度を最大化（ノルムまたは $\cosine$ 距離を最小化）するための第三の学習に相当する。
- [0038] このようにして、本実施の形態に係る学習装置 1 0 は、3つの学習（パラメータ更新）処理を組み合わせることによって、第一言語および第二言語のそれぞれのテキストを一つの意味空間上にベクトルとして圧縮変換することが可能となり、意味が同じテキストであれば、同等の圧縮ベクトルを得られるように学習することができる。
- [0039] また、第三の学習のみでは過学習が起こる可能性が高いが、第一、第二の学習を組み合わせることにより、汎化の学習効果が得られ、過学習が抑制されることが期待できる。
- [0040] なお、学習装置 1 0 は、学習の組み合わせ方としては、それぞれを一定回数毎に順番に学習を繰り返しても良いし、一定期間の学習誤差を全てプールし、バッチ処理として重みの更新を行っても良い。
- [0041] （対訳文出力装置の機能構成）

図 6 は、対訳文出力装置の機能構成図である。対訳文出力装置 2 0 は、モデル記憶部 2 1 と、第一言語テキスト取得部 2 2 と、第一言語圧縮ベクトル生成部 2 3 と、第二言語対訳文候補取得部 2 4 と、第二言語圧縮ベクトル生成部 2 5 と、類似度算出部 2 6 と、対訳文選択部 2 7 と、出力部 2 8 と、を

備える。

- [0042] モデル記憶部 21 は、言語モデル 200 を記憶する。言語モデル 200 は、学習装置 10 によってパラメータを更新された学習済みの言語モデルである。
- [0043] 第一言語テキスト取得部 22 は、第一言語テキスト 90 を取得する。第一言語テキスト 90 は、例えば日本語を第一言語として対訳文を出力する場合、対訳の対象となる日本語のテキストである。
- [0044] 第一言語圧縮ベクトル生成部 23 は、言語モデル 200 に基づいて第一言語のテキストを圧縮して、圧縮ベクトル（第一圧縮ベクトル）を生成する。
- [0045] 第二言語対訳文候補取得部 24 は、第二言語対訳文候補 91 を取得する。第二言語対訳文候補 91 は、例えば英語を第二言語とする場合、第一言語の一例である日本語の対訳の候補となる英語のテキストの集合である。
- [0046] 第二言語圧縮ベクトル生成部 25 は、言語モデル 200 に基づいて第二言語の対訳文候補を圧縮して、圧縮ベクトル（第二圧縮ベクトル）を生成する。
- [0047] 類似度算出部 26 は、第一言語のテキストから生成された圧縮ベクトル（第一圧縮ベクトル）と、第二言語の複数の対訳文候補から生成されたそれぞれの圧縮ベクトル（第二圧縮ベクトル）との類似度を算出する。類似度の算出方法は、学習装置 10 の類似度算出部 18 と同様である。
- [0048] 対訳文選択部 27 は、複数の対訳文の候補から、類似度に基づいて対訳文を選択する。具体的には、類似度が圧縮ベクトルの $\cos$ 距離である場合、対訳文選択部 27 は、対訳文候補から生成された複数の圧縮ベクトルのうち、 $\cos$ 距離が最小となる圧縮ベクトルの生成元の対訳文を選択する。
- [0049] 出力部 28 は、選択された対訳文を出力する。具体的には、出力部 28 は、対訳文を示す情報を他の装置等に送信するか、または表示装置等に表示する。
- [0050] （対訳文出力処理における言語モデル）

図 7 は、対訳文出力処理における言語モデルの使用方法について説明するための図である。図 7 に示す言語モデル 200 は、図 2 に示した言語モデル 100 を対訳文の出力処理に適用したものである。

[0051] 言語モデル 200 は、(1) 第一言語のテキストの圧縮ベクトルの生成に関するエンコーダ 201 および中間層 202 と、(2) 第二言語の対訳文候補の圧縮ベクトルの生成に関するエンコーダ 203 および中間層 204 と、(3) 類似度の算出に関する中間層 205 と、(4) 算出された類似度を示す情報を記憶する記憶部 206 と、を含む。

[0052] (対訳文出力装置の動作例)

次に、対訳文出力装置 20 の動作例について、図面を参照して説明する。対訳文出力装置 20 は、ユーザの操作等を受けて、または定期的に、対訳文出力処理を開始する。

[0053] 図 8 は、対訳文出力処理の流れの一例を示すフローチャートである。第一言語テキスト取得部 22 は、第一言語テキストを取得する(ステップ S 21)。次に、第一言語圧縮ベクトル生成部 23 は、第一言語テキストの圧縮ベクトルを生成する(ステップ S 22)。

[0054] 続いて、第二言語対訳文候補取得部 24 は、第二言語の対訳文の候補を取得する(ステップ S 23)。第二言語圧縮ベクトル生成部 25 は、第二言語の対訳文の候補の圧縮ベクトルを生成する(ステップ S 24)。

[0055] 次に、類似度算出部 26 は、第一言語テキストの圧縮ベクトルと、第二言語の対訳文の候補の圧縮ベクトルとの類似度を算出する(ステップ S 25)。対訳文選択部 27 は、類似度に基づいて対訳文を選択する(ステップ S 26)。

[0056] そして、出力部 28 は、選択された対訳文を出力する(ステップ S 27)。

[0057] (本実施の形態に係るハードウェア構成例)

学習装置 10 および対訳文出力装置 20 は、例えば、コンピュータに、本実施の形態で説明する処理内容を記述したプログラムを実行させることによ

り実現可能である。なお、この「コンピュータ」は、物理マシンであってもよいし、クラウド上の仮想マシンであってもよい。仮想マシンを使用する場合、ここで説明する「ハードウェア」は仮想的なハードウェアである。

[0058] 上記プログラムは、コンピュータが読み取り可能な記録媒体（可搬メモリ等）に記録して、保存したり、配布したりすることが可能である。また、上記プログラムをインターネットや電子メール等、ネットワークを通して提供することも可能である。

[0059] 図9は、上記コンピュータのハードウェア構成例を示す図である。図9のコンピュータは、それぞれバスBで相互に接続されているドライブ装置1000、補助記憶装置1002、メモリ装置1003、CPU1004、インタフェース装置1005、表示装置1006、入力装置1007、出力装置1008等を有する。

[0060] 当該コンピュータでの処理を実現するプログラムは、例えば、CD-ROM又はメモリカード等の記録媒体1001によって提供される。プログラムを記憶した記録媒体1001がドライブ装置1000にセットされると、プログラムが記録媒体1001からドライブ装置1000を介して補助記憶装置1002にインストールされる。但し、プログラムのインストールは必ずしも記録媒体1001より行う必要はなく、ネットワークを介して他のコンピュータよりダウンロードするようにしてもよい。補助記憶装置1002は、インストールされたプログラムを格納すると共に、必要なファイルやデータ等を格納する。

[0061] メモリ装置1003は、プログラムの起動指示があった場合に、補助記憶装置1002からプログラムを読み出して格納する。CPU1004は、メモリ装置1003に格納されたプログラムに従って、当該装置に係る機能を実現する。インタフェース装置1005は、ネットワークに接続するためのインタフェースとして用いられる。表示装置1006はプログラムによるGUI（Graphical User Interface）等を表示する。入力装置1007はキーボード及びマウス、ボタン、又はタッチパネル等

で構成され、様々な操作指示を入力させるために用いられる。出力装置 1008 は演算結果を出力する。なお、上記コンピュータは、CPU 1004 の代わりに GPU (Graphics Processing Unit) または TPU (Tensor processing unit) を備えていても良く、CPU 1004 に加えて、GPU または TPU を備えていても良い。その場合、例えばニューラルネットワーク等の特殊な演算が必要な処理を GPU または TPU が実行し、その他の処理を CPU 1004 が実行する、というように処理を分担して実行しても良い。

[0062] (本実施の形態の効果)

本実施の形態に係る学習装置 10 によれば、3つの学習（パラメータ更新）処理を組み合わせることによって、第一言語および第二言語のそれぞれのテキストを一つの意味空間上にベクトルとして圧縮変換することが可能となり、意味が同じテキストであれば、同等の圧縮ベクトルを得られるように学習する。

[0063] 近年ではインターネット等を介して膨大な数の単言語のコーパスが提供されているため、単言語のコーパスの収集は比較的容易である。したがって、入手しやすい単言語のコーパスを利用することによって、大量の学習データを利用でき、圧縮次元における意味空間の滑らかさを実現させることができる。

[0064] このようにして学習された言語モデルを利用することによって、対訳文出力装置 20 は、異なる言語間で意味の近い文を発見することが容易になり、精度の高い対訳文を選択して出力することができる。

[0065] (参考文献)

[1] : The Hugging Face Team, Transformers, [online], インターネット<URL : <https://huggingface.co/transformers/index.html>>

[0066] (実施の形態のまとめ)

本明細書には、少なくとも下記の各項に記載した学習装置、対訳文出力装置、学習方法、対訳文出力方法およびプログラムが記載されている。

(第1項)

言語モデルのパラメータを更新するための学習装置であって、
前記言語モデルに基づいて、第一言語のコーパスを圧縮した第一圧縮ベクトルを生成する第一言語圧縮ベクトル生成部と、
前記第一圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第一パラメータ更新部と、
前記言語モデルに基づいて、第二言語のコーパスを圧縮した第二圧縮ベクトルを生成する第二言語圧縮ベクトル生成部と、
前記第二圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第二パラメータ更新部と、
それぞれ意味が同じテキストから生成された前記第一圧縮ベクトルと前記第二圧縮ベクトルの類似度を算出する類似度算出部と、
前記類似度に基づいて、言語モデルのパラメータを更新する第三パラメータ更新部と、を備える、
学習装置。

(第2項)

前記類似度算出部は、前記第一圧縮ベクトルと前記第二圧縮ベクトルとのノルムまたは $\cos$ ine距離を、前記類似度として算出し、
前記第三パラメータ更新部は、前記ノルムまたは前記 $\cos$ ine距離を最小化するように前記パラメータを更新する、
第1項に記載の学習装置。

(第3項)

前記言語モデルは、第一言語の圧縮ベクトルを中間層に生成するエンコーダと、第二言語の圧縮ベクトルを中間層に生成するエンコーダと、を含むオートエンコーダである、
第1項または第2項に記載の学習装置。

(第4項)

第一言語のテキストの第二言語の対訳文を出力するための対訳文出力装置であって、

前記第一言語のテキストを圧縮して第一圧縮ベクトルを生成する第一言語圧縮ベクトル生成部と、

前記第二言語の複数の対訳文候補を圧縮して第二圧縮ベクトルを生成する第二言語圧縮ベクトル生成部と、

前記第一圧縮ベクトルと、それぞれの前記第二圧縮ベクトルとの類似度を算出する類似度算出部と、

前記類似度に基づいて対訳文候補から対訳文を選択する対訳文選択部と、
選択された前記対訳文を出力する出力部と、を備える、
対訳文出力装置。

(第5項)

言語モデルのパラメータを更新するための学習装置が実行する学習方法であって、

前記言語モデルに基づいて、第一言語のコーパスを圧縮した第一圧縮ベクトルを生成するステップと、

前記第一圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新するステップと、

前記言語モデルに基づいて、第二言語のコーパスを圧縮した第二圧縮ベクトルを生成するステップと、

前記第二圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新するステップと、

それぞれ意味が同じテキストから生成された前記第一圧縮ベクトルと前記第二圧縮ベクトルの類似度を算出するステップと、

前記類似度に基づいて、言語モデルのパラメータを更新するステップと、
を備える、

学習方法。

(第6項)

第一言語のテキストの第二言語の対訳文を出力するための対訳文出力装置が実行する対訳文出力方法であって、

前記第一言語のテキストを圧縮して第一圧縮ベクトルを生成するステップと、

前記第二言語の複数の対訳文候補を圧縮して第二圧縮ベクトルを生成するステップと、

前記第一圧縮ベクトルと、それぞれの前記第二圧縮ベクトルとの類似度を算出するステップと、

前記類似度に基づいて対訳文候補から対訳文を選択するステップと、

選択された前記対訳文を出力するステップと、を備える、

対訳文出力方法。

(第7項)

コンピュータを、第1項から第3項のいずれか1項に記載の学習装置における各部として機能させるためのプログラム、または、コンピュータを、第4項に記載の対訳文出力装置における各部として機能させるためのプログラム。

[0067] 以上、本実施の形態について説明したが、本発明はかかる特定の実施形態に限定されるものではなく、請求の範囲に記載された本発明の要旨の範囲内において、種々の変形・変更が可能である。

符号の説明

- [0068] 10 学習装置
- 11 モデル記憶部
 - 12 第一言語コーパス取得部
 - 13 第一言語圧縮ベクトル生成部
 - 14 第一パラメータ更新部
 - 15 第二言語コーパス取得部
 - 16 第二言語圧縮ベクトル生成部
 - 17 第二パラメータ更新部
 - 18 類似度算出部
 - 19 第三パラメータ更新部

- 2 0 対訳文出力装置
- 2 1 モデル記憶部
- 2 2 第一言語テキスト取得部
- 2 3 第一言語圧縮ベクトル生成部
- 2 4 第二言語対訳文候補取得部
- 2 5 第二言語圧縮ベクトル生成部
- 2 6 類似度算出部
- 2 7 対訳文選択部
- 2 8 出力部
- 8 0 第一言語コーパス
- 8 1 第二言語コーパス
- 9 0 第一言語テキスト
- 9 1 第二言語対訳文候補
- 1 0 0 言語モデル
- 2 0 0 言語モデル
- 1 0 0 0 ドライブ装置
- 1 0 0 1 記録媒体
- 1 0 0 2 補助記憶装置
- 1 0 0 3 メモリ装置
- 1 0 0 4 C P U
- 1 0 0 5 インタフェース装置
- 1 0 0 6 表示装置
- 1 0 0 7 入力装置
- 1 0 0 8 出力装置

請求の範囲

- [請求項1] 言語モデルのパラメータを更新するための学習装置であって、
前記言語モデルに基づいて、第一言語のコーパスを圧縮した第一圧縮ベクトルを生成する第一言語圧縮ベクトル生成部と、
前記第一圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第一パラメータ更新部と、
前記言語モデルに基づいて、第二言語のコーパスを圧縮した第二圧縮ベクトルを生成する第二言語圧縮ベクトル生成部と、
前記第二圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新する第二パラメータ更新部と、
それぞれ意味が同じテキストから生成された前記第一圧縮ベクトルと前記第二圧縮ベクトルの類似度を算出する類似度算出部と、
前記類似度に基づいて、言語モデルのパラメータを更新する第三パラメータ更新部と、を備える、
学習装置。
- [請求項2] 前記類似度算出部は、前記第一圧縮ベクトルと前記第二圧縮ベクトルとのノルムまたは $\cos$ 距離を、前記類似度として算出し、
前記第三パラメータ更新部は、前記ノルムまたは前記 $\cos$ 距離を最小化するように前記パラメータを更新する、
請求項1に記載の学習装置。
- [請求項3] 前記言語モデルは、第一言語の圧縮ベクトルを中間層に生成するエンコーダと、第二言語の圧縮ベクトルを中間層に生成するエンコーダと、を含むオートエンコーダである、
請求項1または2に記載の学習装置。
- [請求項4] 第一言語のテキストの第二言語の対訳文を出力するための対訳文出力装置であって、
前記第一言語のテキストを圧縮して第一圧縮ベクトルを生成する第一言語圧縮ベクトル生成部と、

前記第二言語の複数の対訳文候補を圧縮して第二圧縮ベクトルを生成する第二言語圧縮ベクトル生成部と、

前記第一圧縮ベクトルと、それぞれの前記第二圧縮ベクトルとの類似度を算出する類似度算出部と、

前記類似度に基づいて対訳文候補から対訳文を選択する対訳文選択部と、

選択された前記対訳文を出力する出力部と、を備える、

対訳文出力装置。

[請求項5] 言語モデルのパラメータを更新するための学習装置が実行する学習方法であって、

前記言語モデルに基づいて、第一言語のコーパスを圧縮した第一圧縮ベクトルを生成するステップと、

前記第一圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新するステップと、

前記言語モデルに基づいて、第二言語のコーパスを圧縮した第二圧縮ベクトルを生成するステップと、

前記第二圧縮ベクトルに基づいて、前記言語モデルのパラメータを更新するステップと、

それぞれ意味が同じテキストから生成された前記第一圧縮ベクトルと前記第二圧縮ベクトルの類似度を算出するステップと、

前記類似度に基づいて、言語モデルのパラメータを更新するステップと、を備える、

学習方法。

[請求項6] 第一言語のテキストの第二言語の対訳文を出力するための対訳文出力装置が実行する対訳文出力方法であって、

前記第一言語のテキストを圧縮して第一圧縮ベクトルを生成するステップと、

前記第二言語の複数の対訳文候補を圧縮して第二圧縮ベクトルを生

成するステップと、

前記第一圧縮ベクトルと、それぞれの前記第二圧縮ベクトルとの類似度を算出するステップと、

前記類似度に基づいて対訳文候補から対訳文を選択するステップと

、

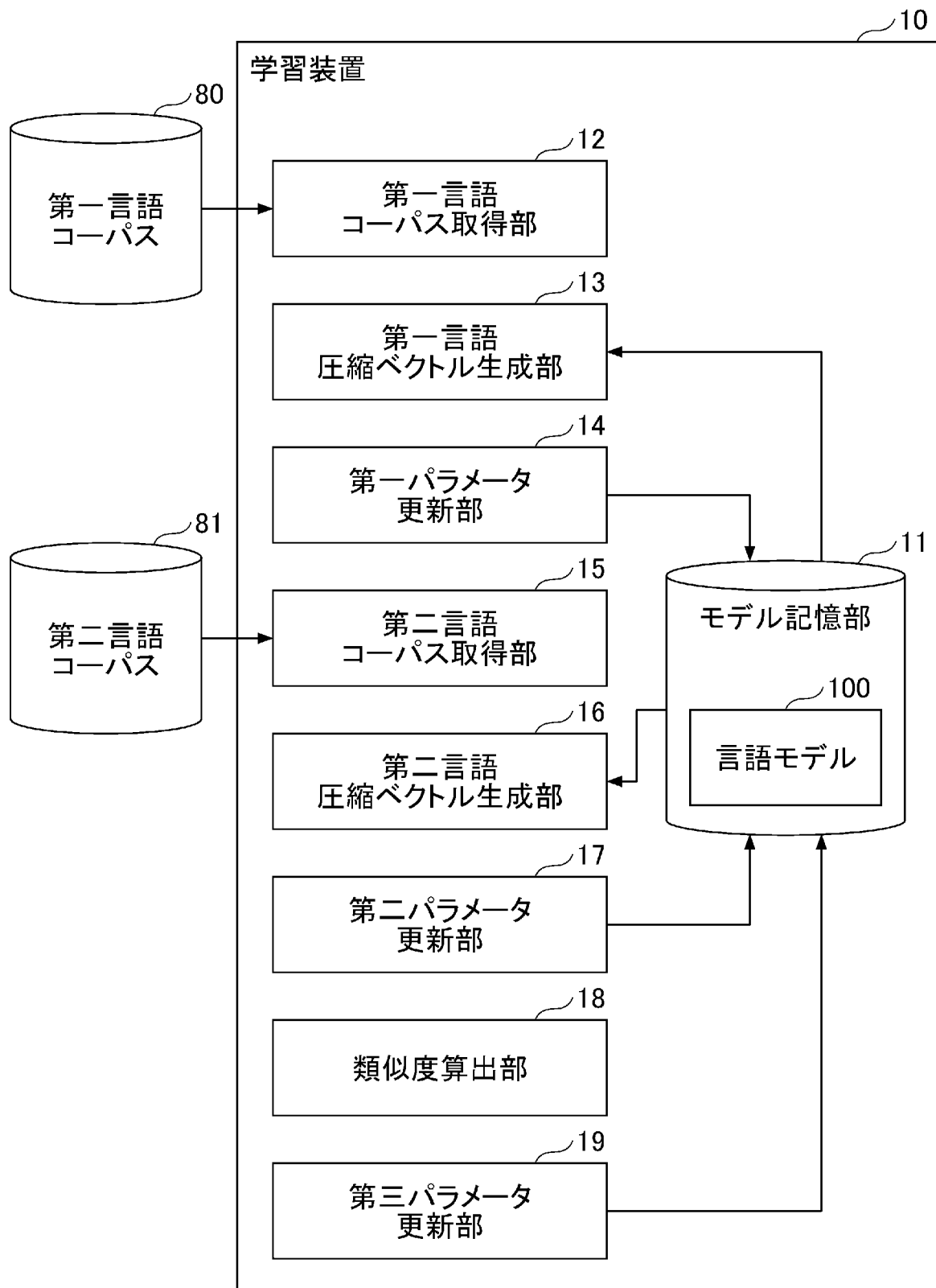
選択された前記対訳文を出力するステップと、を備える、

対訳文出力方法。

[請求項7]

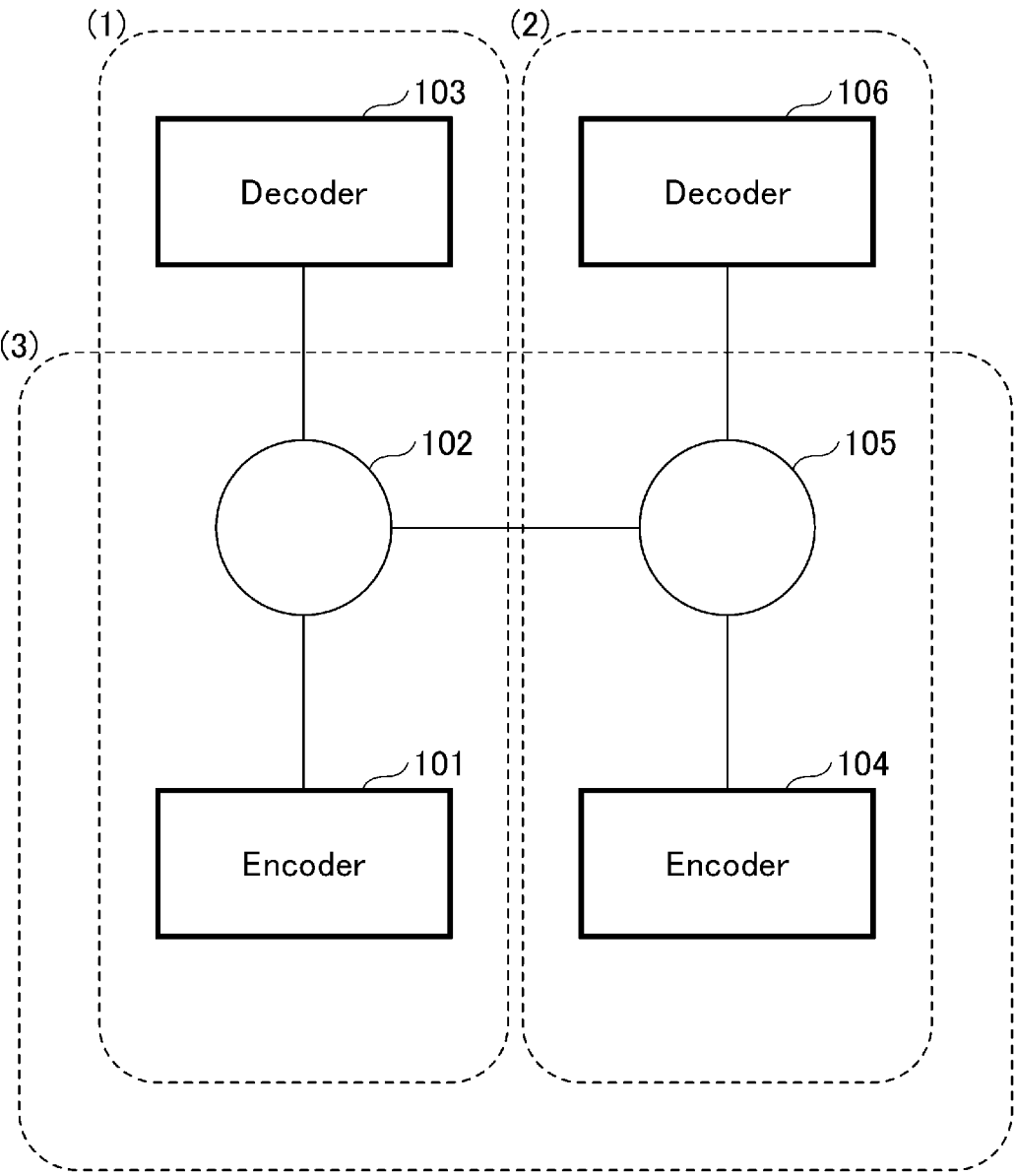
コンピュータを、請求項1から3のいずれか1項に記載の学習装置における各部として機能させるためのプログラム、または、コンピュータを、請求項4に記載の対訳文出力装置における各部として機能させるためのプログラム。

[図1]



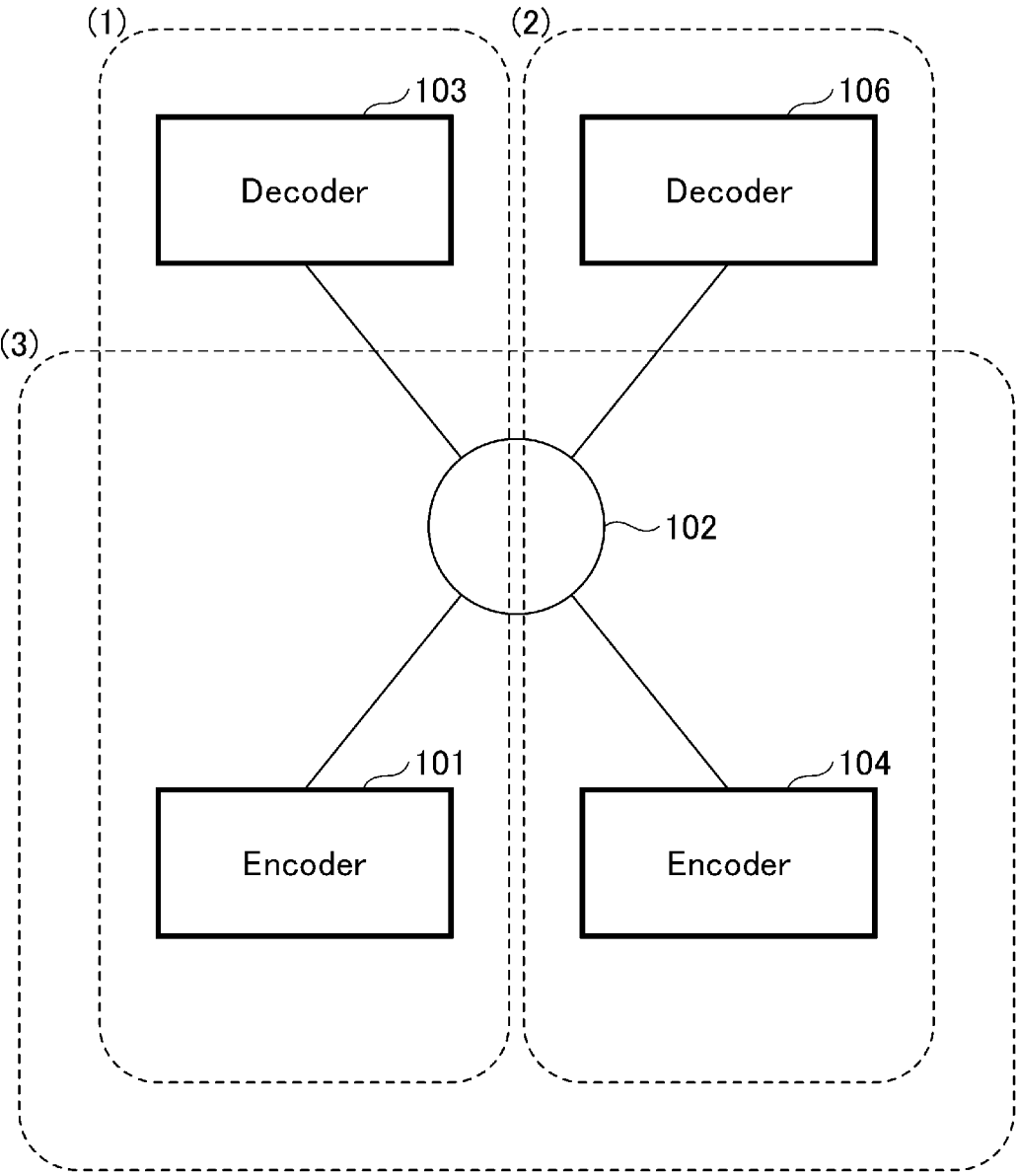
[図2]

100

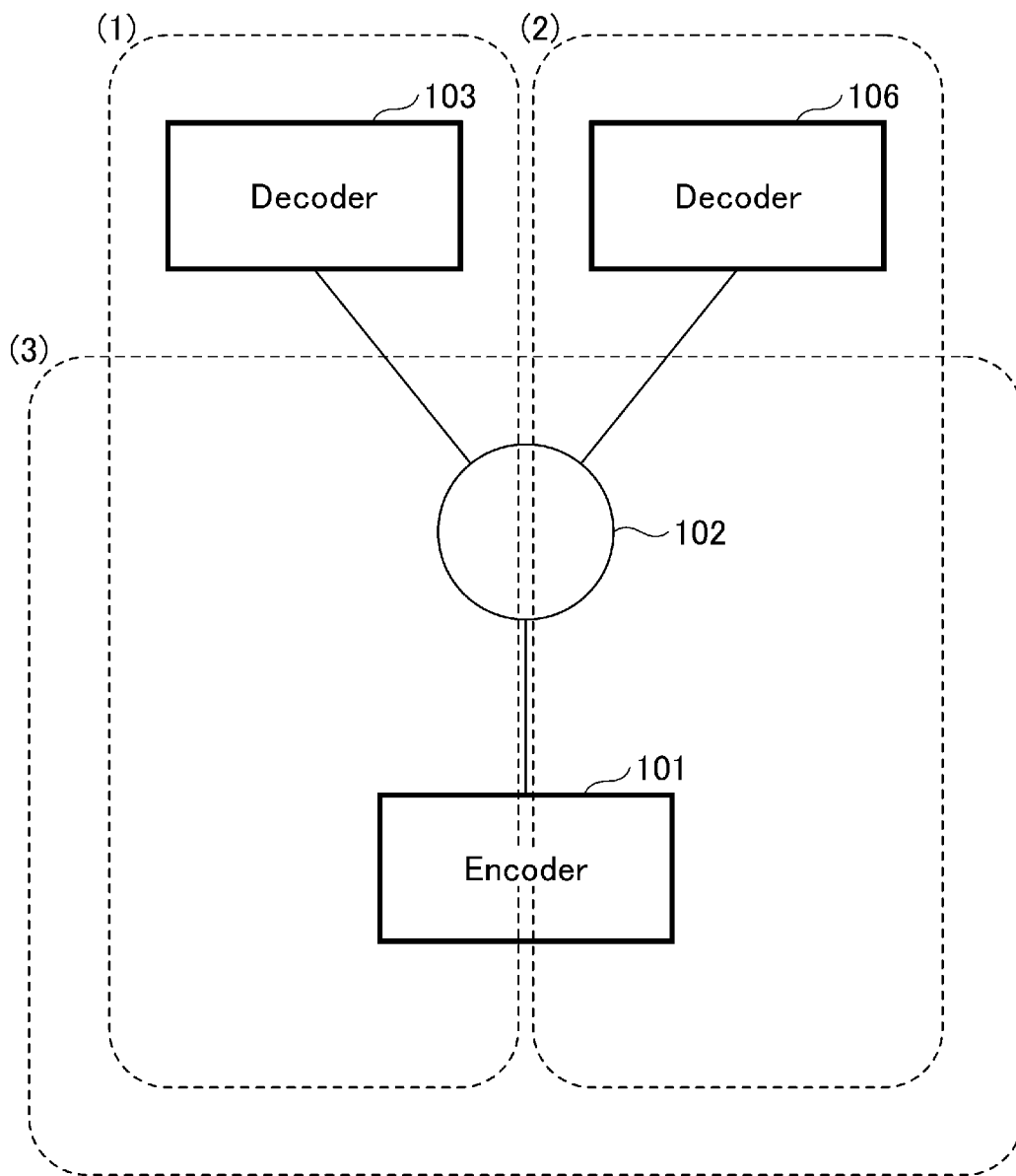


[図3]

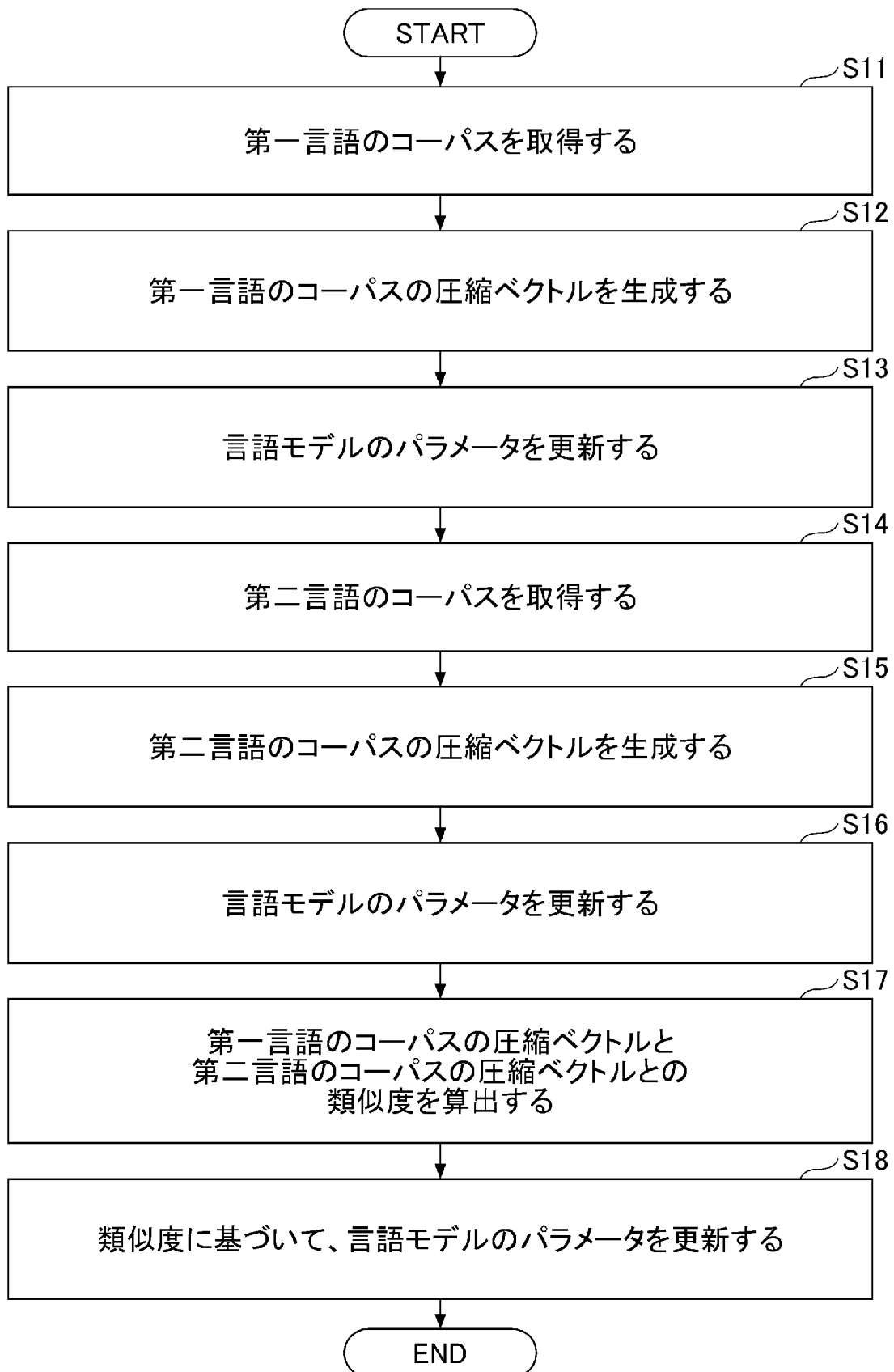
100



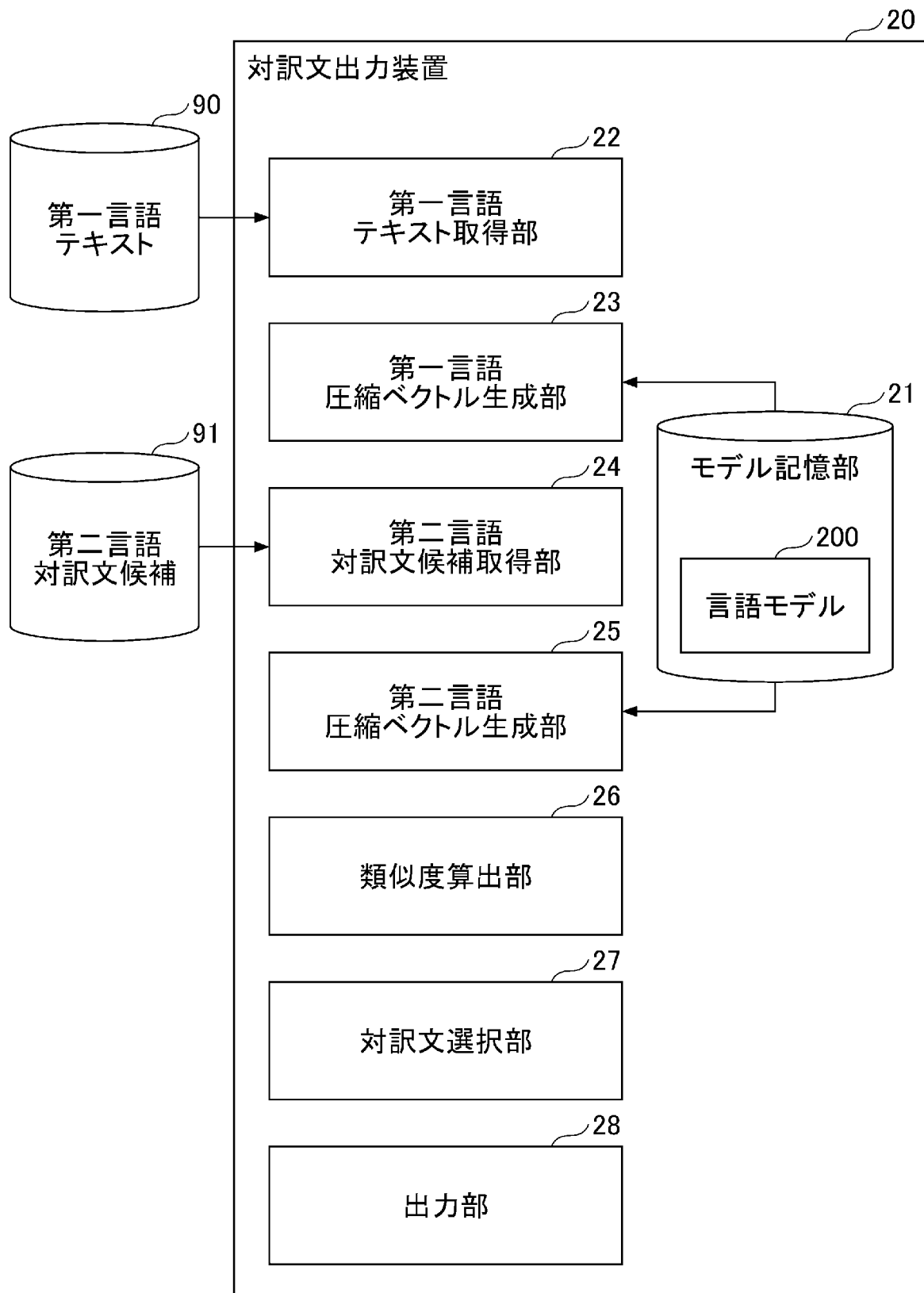
[図4]

100

[図5]

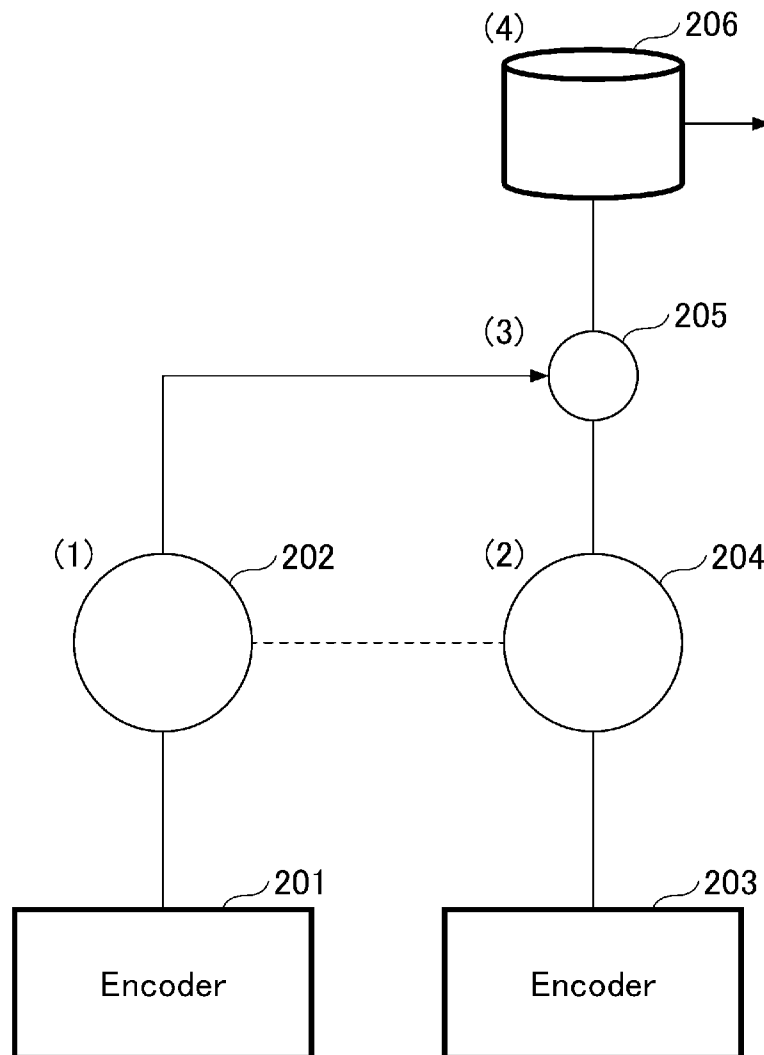


[図6]

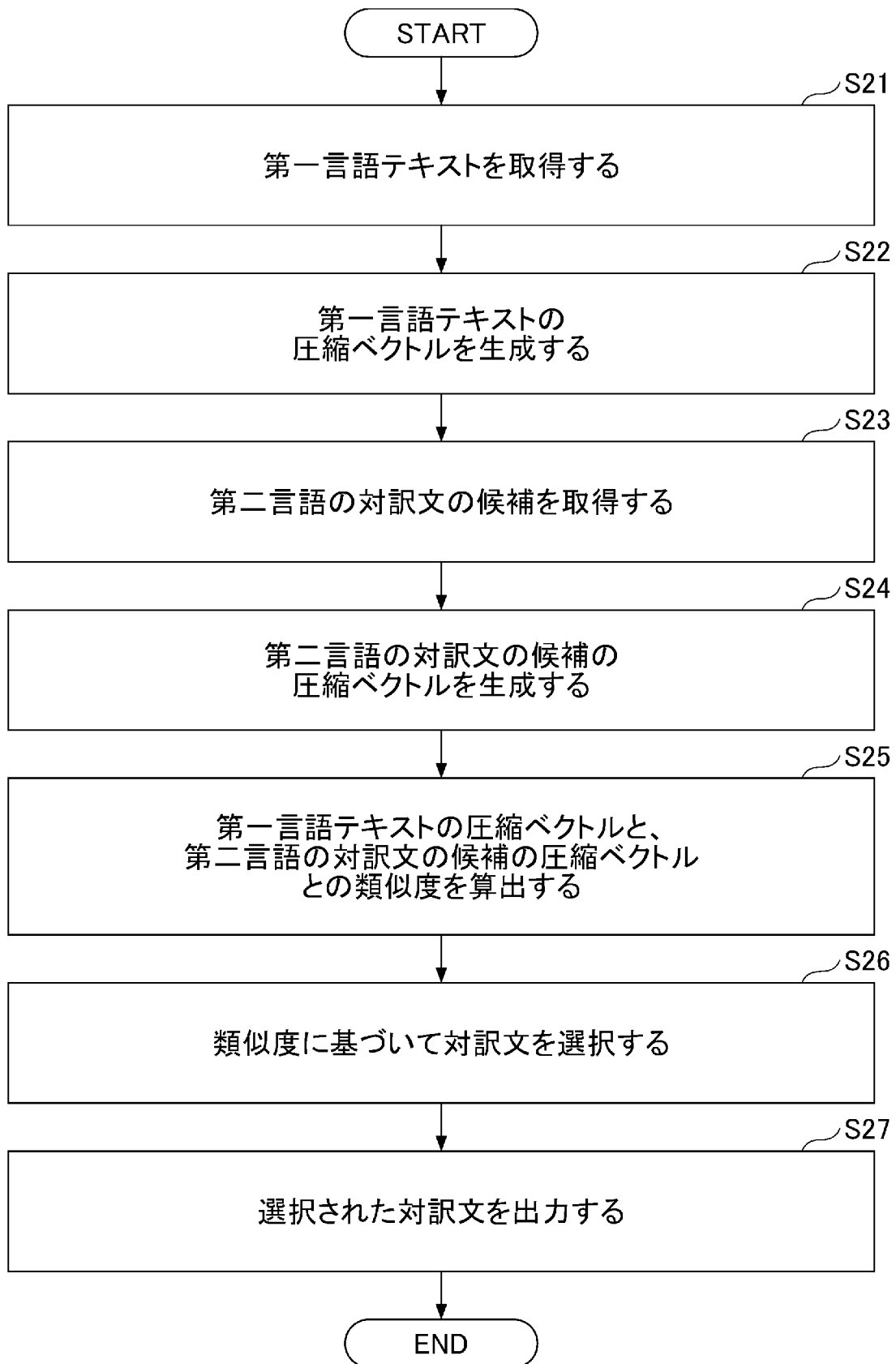


[図7]

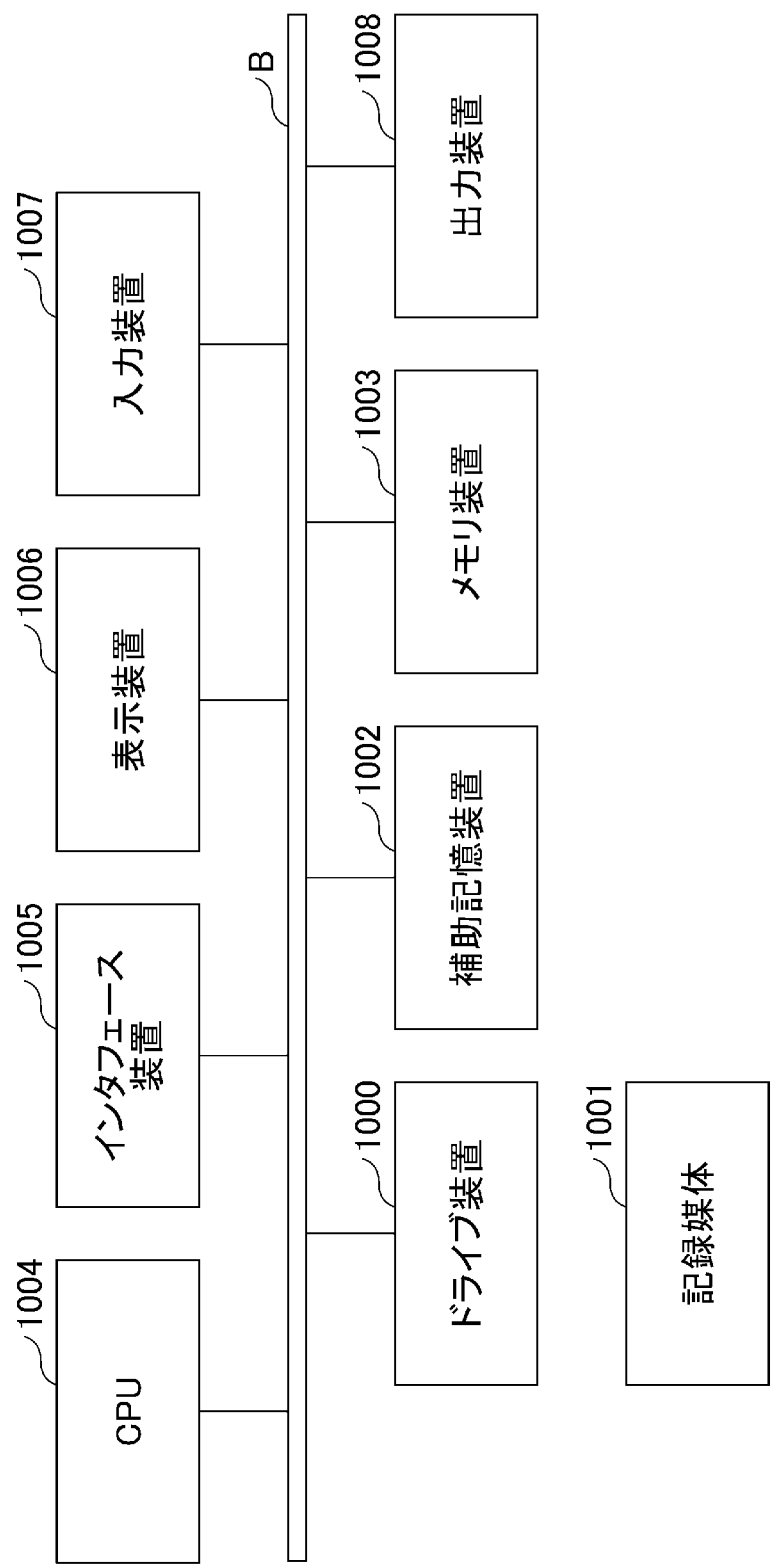
200



[図8]



[図9]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2021/022581

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/44(2020.01)i; G06F 40/45(2020.01)i
FI: G06F40/44; G06F40/45

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)

G06F40/20-40/58; G06N3/00-99/00

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	1971-2021
Registered utility model specifications of Japan	1996-2021
Published registered utility model applications of Japan	1994-2021

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.
X	US 2017/0060854 A1 (ALIBABA GROUP HOLDING LTD) 02 March 2017 (2017-03-02) claims 1-7	1-7
A	WO 2018/055983 A1 (PANASONIC IP MAN CO LTD) 29 March 2018 (2018-03-29) paragraphs [0076]-[0094]	1-7
A	JP 2014-10634 A (NATIONAL INSTITUTE OF INFORMATION & COMMUNICATIONTECHNOLOGY) 20 January 2014 (2014-01-20) entire text, all drawings	1-7

☐ 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
11 August 2021 (11.08.2021)

Date of mailing of the international search report
24 August 2021 (24.08.2021)

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/JP2021/022581

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US 2017/0060854 A1	02 Mar. 2017	WO 2017/033063 A2 CN 106484682 A TW 201717070 A	
WO 2018/055983 A1	29 Mar. 2018	US 2019/0179908 A1 paragraphs [0092]- [0107]	
JP 2014-10634 A	20 Jan. 2014	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 40/44(2020.01)i; G06F 40/45(2020.01)i FI: G06F40/44; G06F40/45														
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）） G06F40/20-40/58; G06N3/00-99/00														
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														
<table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2021年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2021年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2021年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2021年	日本国実用新案登録公報	1996 - 2021年	日本国登録実用新案公報	1994 - 2021年				
日本国実用新案公報	1922 - 1996年													
日本国公開実用新案公報	1971 - 2021年													
日本国実用新案登録公報	1996 - 2021年													
日本国登録実用新案公報	1994 - 2021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												
X	US 2017/0060854 A1 (ALIBABA GROUP HOLDING LTD) 02.03.2017 (2017 - 03 - 02) 請求項1-7	1-7												
A	WO 2018/055983 A1 (パナソニックIPマネジメント株式会社) 29.03.2018 (2018 - 03 - 29) 段落76-94	1-7												
A	JP 2014-10634 A (独立行政法人情報通信研究機構) 20.01.2014 (2014 - 01 - 20) 全文・全図	1-7												
☐ 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。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>“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</td> </tr> <tr> <td>“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</td> <td>“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</td> </tr> <tr> <td>“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</td> <td>“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</td> </tr> <tr> <td>“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）</td> <td>“&” 同一パテントファミリー文献</td> </tr> <tr> <td>“O” 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>“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</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	“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	“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	“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	“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	“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	“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）	“&” 同一パテントファミリー文献	“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	
* 引用文献のカテゴリー	“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													
“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	“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													
“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	“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													
“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）	“&” 同一パテントファミリー文献													
“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														
国際調査を完了した日 11.08.2021	国際調査報告の発送日 24.08.2021													
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） 成瀬 博之 5N 9192 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/JP2021/022581

引用文献			公表日	パテントファミリー文献			公表日
US	2017/0060854	A1	02.03.2017	WO	2017/033063	A2	
				CN	106484682	A	
				TW	201717070	A	
WO	2018/055983	A1	29.03.2018	US	2019/0179908	A1	
				段落92-107			
JP	2014-10634	A	20.01.2014	(ファミリーなし)			