

KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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, CV, 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, ME, 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：

文生成学習装置、文生成装置、文生成学習方法及びプログラム

技術分野

[0001] 本発明は、文生成学習装置、文生成装置、文生成学習方法及びプログラムに関する。

背景技術

[0002] 質問テキストと、テキストを含む文書画像とを入力として、当該画像の内容に基づいて当該質問に対する回答テキストを生成する質問応答技術がある（例えば、非特許文献1、非特許文献2）。これらの従来技術では、学習データを{文書画像、質問文、回答文}の3つ組とし、まず、文書画像から、テキスト抽出及び文書領域（Title、Paragraph、Captionなど）が抽出され、テキストと文書画像における文書領域の座標・視覚情報（視覚によって確認できる情報）とが獲得される。そして、獲得された情報及び与えられた質問文を入力とし、回答の生成を行うように学習が行われる。

先行技術文献

非特許文献

[0003] 非特許文献1：Powalski Rafal', Borchmann L'ukasz (l'は、lにクレスカを示し、L'は、Lにクレスカを示す。), Jurkiewicz Dawid, Dwojak Tomasz, Pietruszka Michal', Pal'ka Gabriela: Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. ICDAR21

非特許文献2：Ryota Tanaka, Kyosuke Nishida, Sen Yoshida: VisualMRC: Machine Reading Comprehension on Document Images. AAAI21

発明の概要

発明が解決しようとする課題

[0004] 例えば、「2013年におけるN社の携帯会社におけるモバイルビジネス

シェアはどのくらいですか？」という質問を考える。また、「N社の携帯会社」が「ND社」であることを理解するために必要なページAに対応する文書画像と、「2013年における様々な企業のモバイルビジネスシェア」が記載された、ページAとは別のページBに対応する文書画像があるとする。この例の場合、どちらか一方のページ（文書画像）を参照しても、質問文を理解すること（つまり、質問文に対する回答文を導出すること）は困難であり、これらのページ間（文書画像間）の関係性を考慮する必要がある。しかし、従来技術では、一枚の文書画像が入力されることを想定されており、複数枚の文書画像集合が入力することは考慮されていない。

[0005] 本発明は、上記の点に鑑みてなされたものであって、複数の文書画像に基づく文の生成を可能とすることを目的とする。

課題を解決するための手段

[0006] そこで上記課題を解決するため、文生成学習装置は、複数の文書画像のそれぞれが含む情報と、前記情報に基づく文の生成に関する制約情報と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習するように構成されている学習部、を有する。

発明の効果

[0007] 複数の文書画像に基づく文の生成を可能とすることができる。

図面の簡単な説明

[0008] [図1]本発明の実施の形態における文生成装置10のハードウェア構成例を示す図である。

[図2]本発明の実施の形態における学習時の文生成装置10（文生成学習装置）の機能構成例を示す図である。

[図3]文書画像集合、質問文、正解文の一例を示す図である。

[図4]学習時において文生成装置10が実行する処理手順の一例を説明するためのフローチャートである。

[図5]学習時の文生成モデルm1を説明するための図である。

[図6]本発明の実施の形態における推論時の文生成装置10の機能構成例を示す図である。

[図7]推論時において文生成装置10が実行する処理手順の一例を説明するためのフローチャートである。

[図8]推論時の文生成モデルm1を説明するための図である。

発明を実施するための形態

[0009] 以下、図面に基づいて本発明の実施の形態を説明する。図1は、本発明の実施の形態における文生成装置10のハードウェア構成例を示す図である。

図1の文生成装置10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100、補助記憶装置102、メモリ装置103、プロセッサ104、及びインタフェース装置105等を有する。

[0010] 文生成装置10での処理を実現するプログラムは、CD-ROM等の記録媒体101によって提供される。プログラムを記憶した記録媒体101がドライブ装置100にセットされると、プログラムが記録媒体101からドライブ装置100を介して補助記憶装置102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101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102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。

[0011] メモリ装置103は、プログラムの起動指示があった場合に、補助記憶装置102からプログラムを読み出して格納する。プロセッサ104は、CPU若しくはGPU (Graphics Processing Unit)、又はCPU及びGPUであり、メモリ装置103に格納されたプログラムに従って文生成装置10に係る機能を実行する。インタフェース装置105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。

[0012] 図2は、本発明の実施の形態における学習時の文生成装置10（文生成学習装置）の機能構成例を示す図である。図2において、文生成装置10は、テキスト抽出部11、物体抽出部12、学習部13及び文生成モデルm1等

を有する。これら各部及び文生成モデル $m1$ は、文生成装置10にインストールされた1以上のプログラムが、プロセッサ104に実行させる処理により実現される。文生成装置10は、また、学習データ記憶部121を利用する。学習データ記憶部121は、例えば、補助記憶装置102、又は文生成装置10にネットワークを介して接続可能な記憶装置等を用いて実現可能である。

[0013] 学習データ記憶部121は、文生成モデル $m1$ を学習するための複数の学習データを記憶する。1つの学習データは、質問文、文書画像集合、正解文（回答文又は正解回答根拠ページ番号系列）、タスク名、ページ番号系列の組である。質問文とは、質問を示す文字列（テキスト）である。文書画像集合とは、複数の文書画像の集合である。文書画像とは、テキストや物体の図形又は画像を含む画像である。正解文とは、質問文に対して正解である回答を示す文字列であり、当該文字列の内容はタスクに応じて正解回答文又は正解回答根拠ページ番号系列である。正解回答文とは、質問文に対する回答を示す文である。正解回答根拠ページ番号系列とは、文書画像集合において、正解回答文の根拠となった1以上のページ（文書画像）の識別番号の系列である。タスク名は、タスクの種別を示す情報（種別情報）である。本実施の形態では、タスク名は、「Question Answering」又は「Evidence Selection」である。「Question Answering」は、質問回答タスクのタスク名であり、「Evidence Selection」は、回答根拠選択タスクのタスク名である。但し、これら以外のタスク名が用いられてもよい。質問回答タスクとは、回答文を生成するタスクである。回答根拠選択タスクとは、回答根拠ページ番号系列を選択（又は判定）するタスクである。ページ番号系列は、文書画像集合を構成する各文書画像のページ番号の系列である。

[0014] 例えば、質問文、文書画像集合、正解文（正解回答文又は正解回答根拠ページ番号系列）、タスク名及びページ番号系列の組（学習データ）が N 通り用意され、各組に対してタスク名が異なる2つの学習データが用意される。つまり、 $2 \times N$ 個の学習データが用意される。タスク名が「Question Answer

ing」である学習データの正解文は正解回答文であり、タスク名が「Evidence Selection」である学習データの正解文は正解回答根拠ページ番号系列である。

[0015] 図3は、文書画像集合、質問文、正解文の一例を示す図である。図3には、質問文が、「2013年におけるN社の携帯会社におけるモバイルビジネスシェアはどのくらいですか？」であり、タスク名が「Question Answering」である場合の正解文（＝正解回答文）と、タスク名が「Evidence Selection」である場合の正解文（正解回答根拠ページ番号系列）の一例が示されている。図3の例では、正解回答文である「45%」は、P4及びP5の文書画像が含む情報から生成されるため、正解回答根拠ページ系列は、「4, 5」とされている。

[0016] テキスト抽出部11は、文書画像集合に含まれる文書画像ごとに、当該文書画像において単語の系列を含む領域を検出し、検出された領域内の単語を認識（抽出）して、文書画像中テキスト情報を出力する。文書画像中テキスト情報とは、文書テキスト領域特徴 v_{ocr} 、文書テキスト系列、テキストごとの文書テキスト矩形領域等を含む情報である。

[0017] 文書テキスト矩形領域は、文書画像に対する文字認識によって認識された単語ごとの矩形領域の左上座標と右下座標である。

[0018] 文書テキスト系列は、文書画像において認識された（文書画像から抽出された）単語の系列である。文書テキスト系列は、文書画像において各単語が認識された文書テキスト矩形領域の座標に応じて各テキストがleft-to-right、top-to-downの順番で並び替えられていた単語の系列である。文書テキスト系列及び文書テキスト矩形領域の取得方法は、文書画像から単語系列、及び、単語の矩形領域を出力できる方法であればどのような方法が採用されてもよい。例えば、参考文献1のGoogle（登録商標）Vision APIが用いられてもよい。なお、各参考文献の文献情報は後述される。

[0019] 文書テキスト領域特徴 v_{ocr} は、文書テキスト矩形領域ごとのベクトルの系

列である。1つのベクトルは、2048次元によって1つの文書テキスト矩形領域内の単語の特徴を表す。したがって、文書テキスト領域特徴 v_{ocr} は、 $2048 \times$ 文書テキスト矩形領域数のベクトル系列となる。文書テキスト領域特徴 v_{ocr} の取得にはどのような方法が用いられてもよい。例えば、参考文献2に記載のFaster-RCNNが用いられてもよい。

[0020] 物体抽出部12は、文書画像集合に含まれる文書画像ごとに、当該文書画像に含まれる物体を検出し、当該物体に関する情報（以下、「文書画像中物体情報」という。）を出力する。文書画像中物体情報とは、物体ごとの物体領域特徴 v_{obj} 、物体ごとの物体領域意味ラベル、物体ごとの物体矩形領域等を含む情報である。

[0021] 物体矩形領域は、文書画像に対する物体検出によって認識された物体を含む領域の左上座標及び右下座標である。例えば、物体としては、文書画像中のタイトル、又は表等の図形が物体として検出される。なお、タイトルは、1以上の単語を含む。したがって、タイトルからは、文書画像中テキスト情報及び文書画像中物体情報の双方が抽出される。すなわち、文書画像中の同じ部分に対して、文書画像中テキスト情報及び文書画像中物体情報の双方が抽出される場合がある。

[0022] 物体領域特徴 v_{obj} とは、物体矩形領域ごとのベクトルの系列である。1つのベクトルは、2048次元によって1つの物体矩形領域内の特徴を表す。したがって、物体領域特徴 v_{obj} は、 $2048 \times$ 物体矩形領域数のベクトル系列となる

物体領域意味ラベルは、物体矩形領域の内容（つまり、物体）の意味（名称）を表す9種類の意味ラベル（「Title」や「Table」など）である。「Title」はタイトルの意味ラベルであり、「Table」は表の意味ラベルである。

[0023] 物体領域特徴 v_{obj} 、物体領域意味ラベル、物体矩形領域を文書画像から取得可能であれば文書画像中物体情報を取得するための処理は所定のものに限定されない。例えば、参考文献2のFaster-RCNNが用いられてもよい。

- [0024] 学習部13は、文生成モデルm1の学習を制御する。文生成モデルは、エンコーダとしてのエンコード部mEと、デコーダとしてのデコード部mDとを含む。文生成モデルm1の詳細については後述される。
- [0025] 図4は、学習時において文生成装置10が実行する処理手順の一例を説明するためのフローチャートである。
- [0026] ステップS101において、学習部13は、学習データ記憶部121に記憶されている学習データ群のうちの未処理の一つの学習データ（以下、「対象学習データ」という。）を処理対象として選択する。
- [0027] 続いて、テキスト抽出部11は、対象学習データの文書画像集合に含まれる文書画像ごとに、当該文書画像においてテキストを含む領域を検出し、検出された領域内のテキストを認識して、文書画像中テキスト情報を出力する（S102）。
- [0028] 続いて、物体抽出部12は、対象学習データの文書画像集合に含まれる文書画像ごとに、当該文書画像に含まれる物体を検出し、物体ごとの文書画像中物体情報を出力する（S103）。
- [0029] 続いて、学習部13が、文生成モデルm1のエンコード部mEに対して、ステップS102において抽出された各文書画像の文書画像中テキスト情報、ステップS103において抽出された各文書画像の文書画像中物体情報、対象学習データの質問文、タスク名及び文書画像集合のページ数文のページ番号（つまり、各文書画像のページ番号）の系列であるページ番号系列を入力すると、エンコード部mEは、エンコード特徴量を生成する（S104）。
- [0030] 図5は、学習時の文生成モデルm1を説明するための図である。図5に示されるように、エンコード部mEは、物体特徴受付部mE-1、テキスト情報受付部mE-2、質問文等受付部mE-3及び埋め込み系列生成部mE-4と、M層の変換部層（Transformer）と含む。物体特徴受付部mE-1は、対象学習データの文書画像集合に含まれる各文書画像の文書画像中物体情報（物体領域特徴 v^{obj} 、物体領域意味ラベル、物体矩形領域）の入力を受け付

ける。テキスト情報受付部 $mE-2$ は、各文書画像の文書画像中テキスト情報（文書テキスト領域特徴 v_{ocr} 、文書テキスト系列、文書テキスト矩形領域）の入力を受け付ける。質問文等受付部 $mE-3$ は、対象学習データのページ番号系列、対象学習データの質問文及びタスク名を受け付ける。

[0031] 入力を受け付けに応じ、エンコード部 mE は、以下の $Step1\sim4$ を実行する。

[0032] ($Step1$) テキスト情報受付部 $mE-2$ は、受け付けた各文書テキスト系列を所定の処理単位であるトークンの系列 w_{ocr} に分割する。質問文等受付部 $mE-3$ は、受け付けた質問文を所定の処理単位であるトークンの系列 w_q に分割する。物体特徴受付部 $mE-1$ は、受け付けた各物体領域意味ラベルを所定の処理単位であるトークンの系列 w_{obj} に分割する。所定の処理単位（トークン）への分割処理の一例として、参考文献3に示されるByte-level BPEが用いられてもよい。なお、($Step1$)は、エンコード部 mE の外部、例えばテキスト抽出部11や物体抽出部12で予め実施されてもよい。 k 番目の文書画像について生成されるトークン系列（以下、「文書画像トークン系列」という。） c^k を以下のように定義する。

$$c^k = (w_{1obj}, w_{1ocr}, w_{2obj}, w_{2ocr}, \dots, w_{nobj}, w_{nocr})$$

なお、以下の説明では、 k 番目の文書画像に注目するが、各文書画像について同様の処理が実行される。

[0033] ($Step2$)

埋め込み系列生成部 $mE-4$ は、対象学習データのタスク名 $t \in \{\text{Question Answering, Evidence Selection}\}$ 、対象学習データの質問文のトークン系列 w_q 、対象学習データのページ番号系列のうちの k 番目の文書画像のページ番号 p^k 、 k 番目の文書画像の文書画像トークン系列 c^k を用いて、 k 番目の文書画像の入力トークン系列を以下のように用意（生成）する。

$$x^k = (\text{task} : \text{name} \quad \text{question} : w_q \quad \text{page} : p^k \quad \text{context} : c^k)$$

ここで、“task:”、“question:”、“page:”、“context:”は、順番にタスク名name、質問文のトークン系列 w^q 、ページ番号 p^k 、文書画像トークン系列 c^k の先頭に付与されるトークン系列であり、これらを表現するトークン系列であれば他の文字列によって代替されてもよい。また、ページ番号 $page:p^k$ は、 c^k に含めることも可能である。本実施の形態において、 x^k は系列長 $U=1024$ のトークン系列であるとする。

[0034] (Step 3) 埋め込み系列生成部mE-4は、k番目の文書画像における入力埋め込み系列を以下のように生成する。

$$e^k = \text{LayerNorm}(e_{\text{token}}^k + e_{\text{seg}}^k + e_{\text{vis}}^k + e_{\text{loc}}^k)$$

ここで、LayerNormは、例えば、参考文献4にて示される正規化手法である。また、各入力埋め込み系列は $D \times L$ 次元のベクトルである。本実施の形態では、 $D=512$ 、 $L=200$ である。

[0035] e_{token}^k は、入力トークン系列 x^k を、対応する D 次元のベクトルに埋め込む変換である。本実施の形態では、参考文献5により学習済みの埋め込みベクトル($D=512$)を初期値とし、学習パラメータとする。また、他のEncoder-Decoder型の事前学習済み言語モデルのパラメータが用いられてもよい。

[0036] e_{seg}^k は、入力トークン系列 x^k 中のセグメント（各ページ内のトークンがどの物体領域に属しているかを表す情報）に応じた、 D 次元のベクトル埋め込みである。

[0037] e_{vis}^k は、文書テキスト領域特徴 v_{ocr} と物体領域特徴 v_{obj} とを連結した D 次元のベクトルに変換する処理である。

[0038] e_{loc}^k は、k番目の文書画像の文書テキスト矩形領域、物体矩形領域を表す D 次元の埋め込みである。例えば、非特許文献2で用いられる手法が用いられてもよい。

[0039] なお、 e^k を算出するためには、線型結合の対象となる各ベクトルの系列長を D 次元に揃える必要があるため、 e_{vis}^k 及び e_{loc}^k について、矩形領域を

持たない「タスク名、質問文、ページ番号」に対応する部分はD次元のゼロベクトルにされる。

[0040] (Step 4) 埋め込み系列生成部mE-4は、参考文献6に倣い、各文書画像における入力埋め込み系列 e^k をM層のTransformerに入力する。M層のTransformerは、1つの e^k に対して $e^{k'}$ を生成し、ベクトル連結を行ってエンコード特徴量 $E = [e^{1'}, \dots, e^{K'}]$ を出力する。ここで、Kは文書画像の枚数（ページ数）を表す。

[0041] 本実施の形態では、 $K = 20$ であり、参考文献6で示す事前学習済みの6層のTransformerが用いられる。したがって、本実施の形態において、 $M = 6$ である。但し、他の事前学習済み言語モデルのパラメータが用いられてもよい。

[0042] 上記より、図4のステップS104では、エンコード特徴量 $E = [e^{1'}, \dots, e^{K'}]$ がエンコード部mEから出力される。

[0043] 続いて、文生成モデルm1のデコード部mDは、エンコード特徴量Eと、学習部13から入力される対象学習データのトークナイズ処理済みの正解文（正解回答文又は正解回答根拠ページ番号系列）とに基づいて、出力系列の推定結果である出力系列情報（回答文又は回答根拠ページ番号系列の出力確率分布）を生成する（S105）。すなわち、学習部13は、対象学習データの正解文（正解回答文又は正解回答根拠ページ番号系列）に対してトークナイズ処理を行う。但し、学習部13は、トークナイズ処理済みの正解文を入力としてもよい。例えば、トークナイズ処理として、参考文献3に示されるByte-level BPEが用いられてもよい。

[0044] デコード部mDは、具体的には、以下のStep 11～13を実行する。

[0045] (Step 11)

デコード部mDは、出力ステップt毎に学習部13により入力される正解文（正解回答文又は正解回答根拠ページ番号系列）のトークン系列のうちの $t-1$ 番目のトークンを入力し、t番目のトークンを生成する。そして、デコード部mDは、これまでに出力された各トークンを結合することで出力ス

ステップ t における出力系列を生成する。つまり、学習時において、出力ステップ t における出力系列は、正解文のトークン系列の先頭から t 番目のトークンまでのトークン系列である。但し、本実施の形態においては、回答根拠ページ番号が一つ以上の場合、「 $\langle$ 」を用いてページ番号が連結される。出力ステップとは、1つのトークンがデコード部 mD から出力されるステップをいう。なお、学習部 13 は、正解文のうちデコード部 mD に入力されるトークン系列の先頭に開始記号として " $\langle s \rangle$ Question Answering : "又は " $\langle s \rangle$ Evidence Selection : "を付与し、当該トークン系列の末尾に終端記号として " $\langle /s \rangle$ " 記号を付与する。つまり、" $\langle s \rangle$ [タスク名]" が開始記号として付与される。

[0046] なお、出力系列に対するトークナイズ処理の一例として、参考文献3に示されるByte-level BPEが用いることができる。

[0047] (Step 12)

エンコード部 mE は、エンコード特徴量と出力ステップ t の出力系列を L 層の変換部層 (Transformer) に入力する。その結果、 L 層のTransformerから出力ステップ t の出力トークンの表現 h_t が出力される。本実施の形態においては、参考文献5で示す事前学習済みのTransformerが用いられる。また、同様に、他の事前学習済み言語モデルのパラメータが用いられてもよい。

[0048] (Step 13)

デコード部 mD は、出力トークンの表現 h_t を基に、線形変換と softmax 関数を通すことで、 t 番目のトークンの確率分布 $p(y_t | y_{<t})$ を求める。 t は、 $0 \leq t \leq T$ を満たす。 T は、出力トークン系列の長さの上限であり、本実施の形態においては、 $T = 60$ である。

[0049] (Step 11 ~ 13) が最大で T 回繰り返されると、エンコード部 mE は、出力トークンの表現 $H = [h_1, \dots, h_T]$ に基づく確率分布 p の系列である出力系列情報を出力する。

[0050] 続いて、学習部 13 は、出力系列情報 (出力系列の推定結果) と正解文とに基づいて、以下の損失を算出する (S106)。

[0051] [数1]

$$L = -\frac{1}{T} \sum_t \ln(p(y_t^* | y_{<t}))$$

但し、 y_t^* は、出力テキストにおいて、正解文のトークンである。

[0052] 続いて、学習部13は、当該損失を最小化するよう文生成モデルm1（エンコード部mE及びデコード部mD）のモデルパラメータを更新する（S107）。

[0053] 学習データ記憶部121に記憶されている全ての学習データについてステップS101～S107が実行されると（S108でN）、図4の処理手順は終了する。

[0054] 次に、推論時について説明する。図6は、本発明の実施の形態における推論時の文生成装置10の機能構成例を示す図である。図6中、図2と同一部分には同一符号を付し、その説明は適宜省略する。

[0055] 推論時において、文生成装置10は、出力系列生成部14を有する。出力系列生成部14は、文生成装置10にインストールされた1以上のプログラムが、プロセッサ104に実行させる処理により実現される。

[0056] 出力系列生成部14は、推論時において文生成モデルm1に入力される質問文等について文生成モデルm1が出力する出力系列情報に基づき、回答文又は回答根拠ページ番号系列を生成する。

[0057] なお、推論時において、文生成装置10は、学習部13及び学習データ記憶部121を有さなくてもよい。また、推論時における文生成装置10と学習時における文生成装置10とは異なるコンピュータを用いて実現されてもよい。

[0058] 図7は、推論時において文生成装置10が実行する処理手順の一例を説明するためのフローチャートである。

[0059] ステップS201において、文生成装置10は、入力データの入力を受け

付ける。入力データは、文書画像集合、質問文、タスク名及び文書画像集合を構成する各文書画像のページ番号系列等を含む。

[0060] 続いて、テキスト抽出部11は、入力データの文書画像集合に含まれる文書画像ごとに、当該文書画像に含まれるテキストの領域を検出し、検出された領域内のテキストを認識（抽出）して、文書画像中テキスト情報を出力する（S202）。

[0061] 続いて、物体抽出部12は、入力データの文書画像集合に含まれる文書画像ごとに、当該文書画像に含まれる物体情報を抽出し、文書画像中物体情報を出力する（S203）。

[0062] 続いて、文生成モデルm1のエンコード部mEは、ステップS202において抽出された各文書画像の文書画像中テキスト情報、ステップS203において抽出された各文書画像の文書画像中物体情報、入力データの質問文、タスク名及びページ番号系列を入力して、エンコード特徴量Eを生成する（S204）。

[0063] なお、ステップS202～S204は、処理対象のデータが異なるだけで、処理の内容は図4のステップS102～S104と同じである。続くステップS205～S207は、出力ステップごとに実行される。

[0064] ステップS205において、エンコード部mEは、学習時において説明した（Step11）～（Step13）と同様の処理を1回実行して、t番目のトークンの確率分布 $p(y_t | y_{<t})$ を生成する。

[0065] 但し、推論時において、（Step11）の一部が学習時とは異なる。具体的には、図8に示されるように、推論時においては、出力ステップ毎に出力系列生成部14から出力されるトークンがデコード部mDに入力され、当該トークンを結合したものが出力ステップtにおける出力系列とされる。つまり、推論時において出力ステップtにおける出力系列は、出力系列生成部14がtまでに生成した回答文又は回答根拠ページ系列である。なお、出力系列生成部14は、最初のトークンの先頭に開始記号として“<s>Question Answering:”又は“<s>Evidence Selection:”を付与する。

- [0066] 続いて、出力系列生成部14は、 t 番目のトークンの確率分布 $p(y_t | y_{<t})$ に基づいて、確率分布の最大となるトークンを選択することにより、又は確率分布に従ってトークンをサンプリングすることにより t 番目のトークン y_t （回答文又はページ番号系列の一部）を生成し、トークン y_t を出力する（S206）。
- [0067] トークン y_t が終端記号 $\langle /s \rangle$ でなければ（S207でNo）、ステップS205以降が繰り返される。トークン y_t が終端記号 $\langle /s \rangle$ であれば（S207でYes）、図7の処理手順は終了する。それまでの出力された y_t の系列が、質問文に対する回答文又は回答根拠ページ系列である。
- [0068] 上述したように、本実施の形態によれば、文生成モデル $m1$ は、文書画像集合を入力可能なモデル構造を有する。したがって、複数の文書画像に基づく文の生成を可能とすることができる。その結果、例えば、文書画像間の関係を捉えることが可能になり、実世界に多数存在する視覚的に表現された文書（プレゼンテーションスライド、論文、PDF文書、HTML文書など）を知識源として理解し、文生成（質問応答）を行うことが可能となる。
- [0069] なお、与えられた文書画像集合に対する回答の生成に適合する文書画像と適合しない文書画像との双方が含まれる場合、回答の生成に適合しない文書画像がノイズとなって正しい回答を出力するのが困難となる場合がある。各文書画像が回答の生成に適合するか否か（つまり、どのページが回答の生成に必要であるか否か）の判定は、回答文の生成と密接な関係があるが、従来の技術では、回答の生成に必要なページを選択又は判定するタスクと回答を生成するタスクとを同時に学習することは困難であった。
- [0070] 本実施の形態によれば、質問回答タスクに加えて、回答根拠となるページ番号を選択又は判定するタスクを統一のフォーマット（「Question Answering:」や「Evidence Selection:」といったように「タスク名:」の形でエンコード部 mE に入力し、回答文又は回答根拠ページ番号をトークン系列として生成する、というフォーマット（枠組み）で同時に（単一のモデルで）学習する。その結果、質問文に応じて回答に必要な文書画像（ページ）

を特定し、従来技術よりも高い精度で回答文を生成することが可能になる。

[0071] なお、上記では、文書画像集合に基づく文の生成に関する制約情報が質問文であり、当該質問文に応じた出力系列（回答文又は回答根拠ページ系列）が生成される例について説明したが、他の制約情報を入力とし、当該他の制約情報に応じた出力系列が生成されるタスクに対して本実施の形態が適用されてもよい。

[0072] 例えば、文書画像集合から抽出される情報についての要約文又はキャプション（見出し又はタイトル等）の生成タスクに対して本実施の形態が適用されてもよい。当該タスクに関しては、クエリ（焦点又は着目点を示す文字列）が文生成に関する制約情報として用いられてもよい。

[0073] クエリ指定型要約の場合、{文書画像集合、クエリ、要約文}の三つ組を学習データとして用意することで実現可能である。また、クエリを必要としない抽出型・生成型要約の場合、{文書画像集合、要約文}の二つ組を用意することで実現可能である。

[0074] なお、本実施形態に係る文生成装置10は、非特許文献1や非特許文献2のような従来手法に対して特定の改善を提供するものであり、自然言語による文生成技術が利用される様々な技術分野の向上を示すものである。

[0075] 以下に、本実施の形態において引用した参考文献の文献情報を示す。

[参考文献1] Google Vision API. <https://cloud.google.com/vision>

[参考文献2] Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015: 91-99

[参考文献3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever: Language models are unsupervised multitask learners. Technical report, OpenAI, 2019

[参考文献4] Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey E. Hinton: Layer Normalization. Arxiv, 2016

[参考文献5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee

, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR

[参考文献6] Gautier Izacard and Edouard Grave: Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. EACL2021

以上の実施形態に関し、更に以下の付記を開示する。

[0076] (付記項1)

メモリと、

前記メモリに接続された少なくとも1つのプロセッサと、

を含み、

前記プロセッサは、

複数の文書画像のそれぞれが含む情報と、前記情報に基づく文の生成に関する制約情報と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習する、ことを特徴とする文生成学習装置。

[0077] (付記項2)

メモリと、

前記メモリに接続された少なくとも1つのプロセッサと、

を含み、

前記プロセッサは、

第1の複数の文書画像のそれぞれから抽出された第1の情報と、前記第1の情報に基づく文の生成に関する第1の制約情報と、前記第1の複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて学習された前記モデルが、第2の複数の文書画像のそれぞれから抽出された第2の情報と、前記第2の情報に基づく文の生成に関する第2の制約情報と、

前記第2の複数の文書画像のそれぞれのページ番号とを入力した場合の前記モデルからの出力に基づいて、前記第2の複数の文書画像及び前記第2の制約情報に応じた出力系列を生成する、
ことを特徴とする文生成装置。

[0078] (付記項3)

複数の文書画像のそれぞれが含む情報と、前記情報に基づく文の生成に関する制約情報と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習する、
処理をコンピュータに実行させるプログラムを記録した記録媒体。

[0079] 以上、本発明の実施の形態について詳述したが、本発明は斯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

[0080]	1 0	文生成装置
	1 1	テキスト抽出部
	1 2	物体抽出部
	1 3	学習部
	1 4	出力系列生成部
	1 0 0	ドライブ装置
	1 0 1	記録媒体
	1 0 2	補助記憶装置
	1 0 3	メモリ装置
	1 0 4	プロセッサ
	1 0 5	インタフェース装置
	1 2 1	学習データ記憶部
	B	バス
	m 1	文生成モデル

m E	エンコード部
m E - 1	物体特徴受付部
m E - 2	テキスト情報受付部
m E - 3	質問文等受付部
m E - 4	埋め込み系列生成部
m D	デコード部

請求の範囲

- [請求項1] 複数の文書画像のそれぞれが含む情報と、前記情報に基づく文の生成に関する制約情報と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習するように構成されている学習部、
を有することを特徴とする文生成学習装置。
- [請求項2] 前記モデルは更にタスクの種別情報を入力し、
前記学習部は、
前記モデルが入力したタスクの種別情報が文の生成に関するタスクを示す場合には、前記モデルから出力される、前記制約情報に応じた文の推定結果と、前記入力に対する正解の文とに基づいて前記モデルを学習し、
前記モデルが入力したタスクの種別情報が文の生成に必要なページ番号を判定するタスクを示す場合には、前記モデルから出力される、前記制約情報に応じた文の生成に必要なページ番号の系列と、前記入力に対する正解のページ番号の系列とに基づいて前記モデルを学習するように構成されている、
ことを特徴とする請求項1記載の文生成学習装置。
- [請求項3] 前記制約情報は、質問文であり、
前記学習部は、複数の文書画像のそれぞれから抽出された情報と、前記情報に基づく回答文の生成に関する質問文と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルからの前記回答文に関する出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習するように構成されている、
ことを特徴とする請求項1又は2記載の文生成学習装置。
- [請求項4] 前記モデルは、前記入力に対する特徴量を生成するエンコーダと、前記特徴量に基づいて前記推定結果を生成するデコーダとを含む、

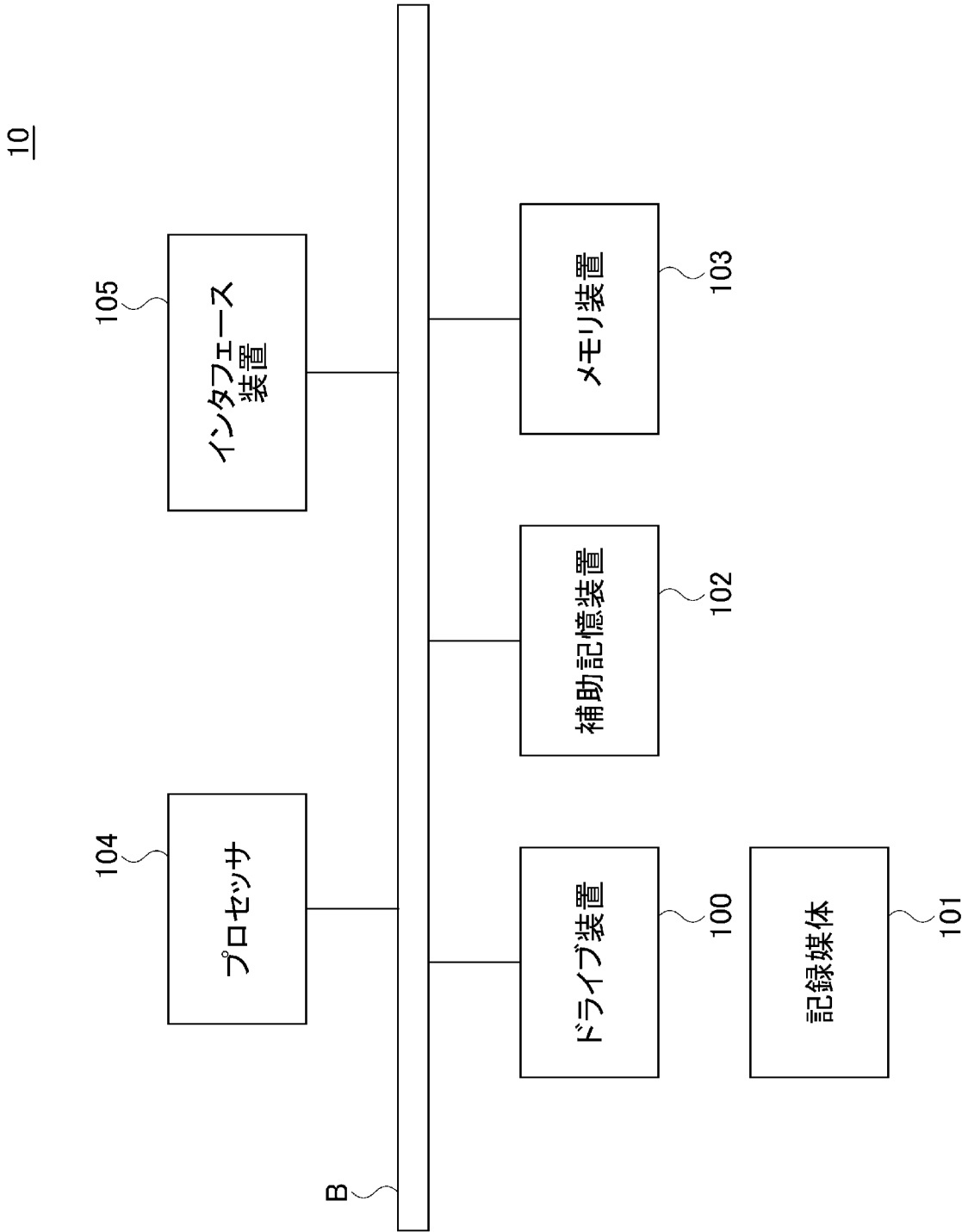
ことを特徴とする請求項 1 又は 2 記載の文生成学習装置。

[請求項5] 第 1 の複数の文書画像のそれぞれから抽出された第 1 の情報と、前記第 1 の情報に基づく文の生成に関する第 1 の制約情報と、前記第 1 の複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて学習された前記モデルが、第 2 の複数の文書画像のそれぞれから抽出された第 2 の情報と、前記第 2 の情報に基づく文の生成に関する第 2 の制約情報と、前記第 2 の複数の文書画像のそれぞれのページ番号とを入力した場合の前記モデルからの出力に基づいて、前記第 2 の複数の文書画像及び前記第 2 の制約情報に応じた出力系列を生成するように構成されている出力系列生成部、を有することを特徴とする文生成装置。

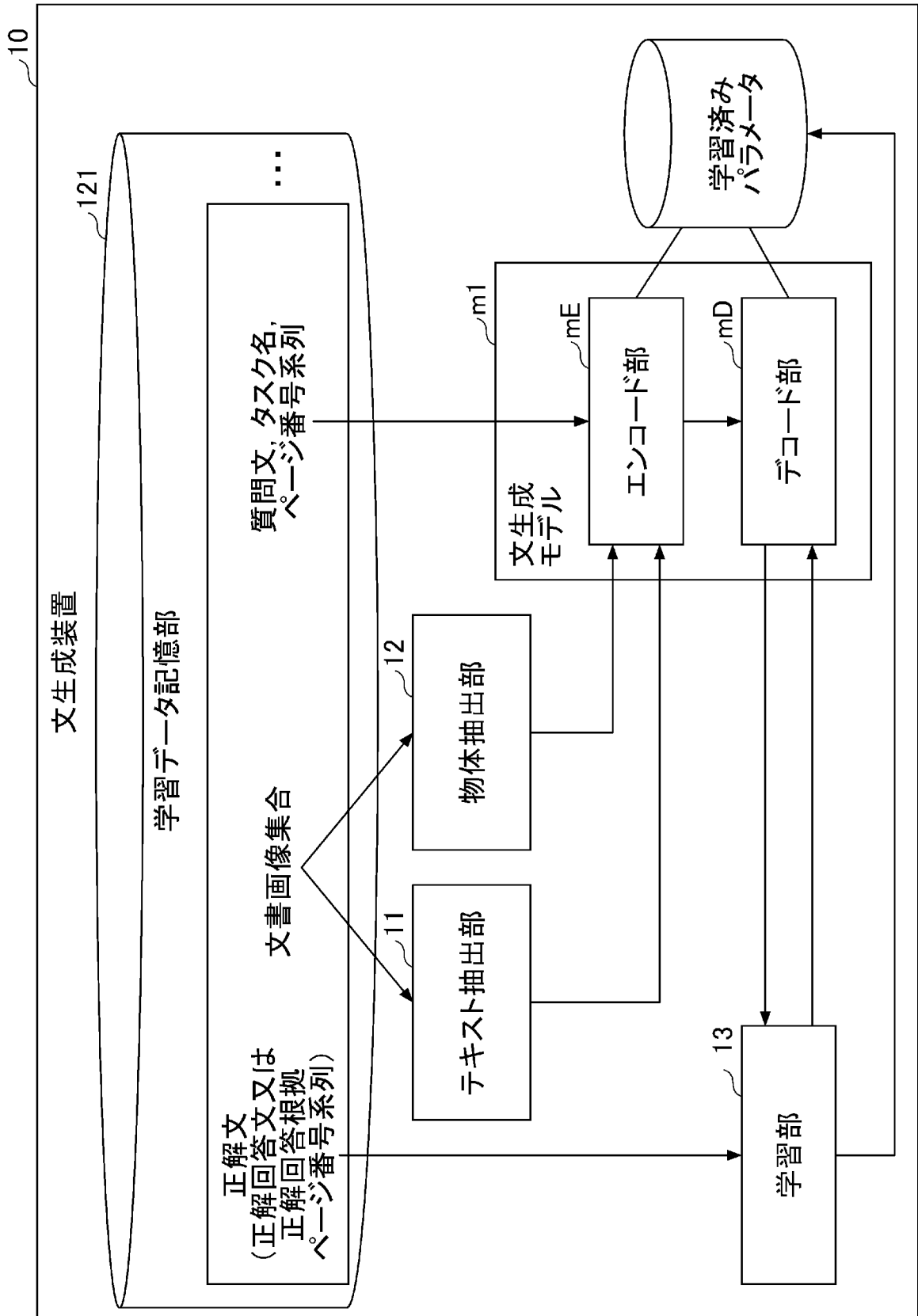
[請求項6] 複数の文書画像のそれぞれが含む情報と、前記情報に基づく文の生成に関する制約情報と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習する学習手順、をコンピュータが実行することを特徴とする文生成学習方法。

[請求項7] 複数の文書画像のそれぞれが含む情報と、前記情報に基づく文の生成に関する制約情報と、前記複数の文書画像のそれぞれのページ番号とを入力として受け付けたモデルから出力される出力系列の推定結果と、前記入力に対する正解の出力系列とに基づいて前記モデルを学習する学習手順、をコンピュータに実行させることを特徴とするプログラム。

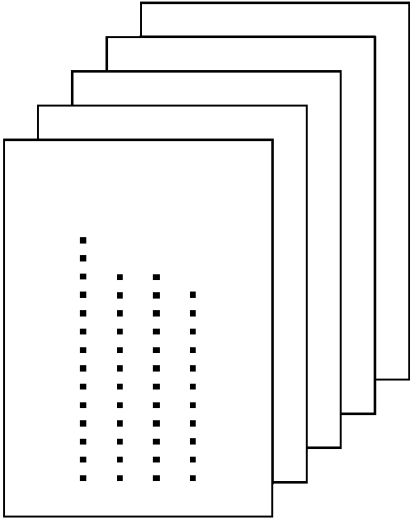
[図1]



[図2]



[図3]

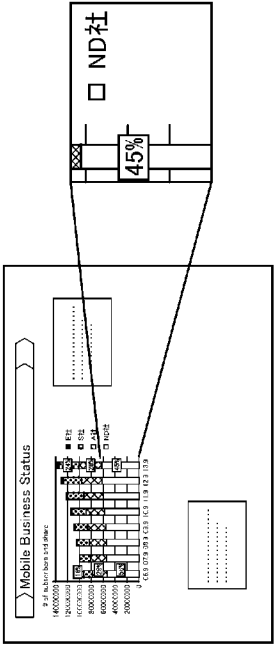


文書画像集合:

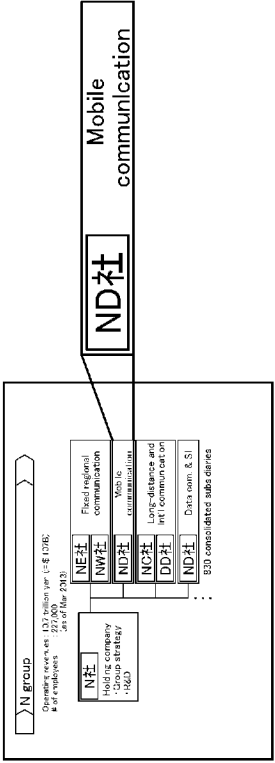
質問文: 2013年におけるN社の携帯電話におけるモバイルビジネスシェアはどのくらいですか?

タスク名=「Question Answering」の場合の正解文(=正解回答文): 45%

p.4

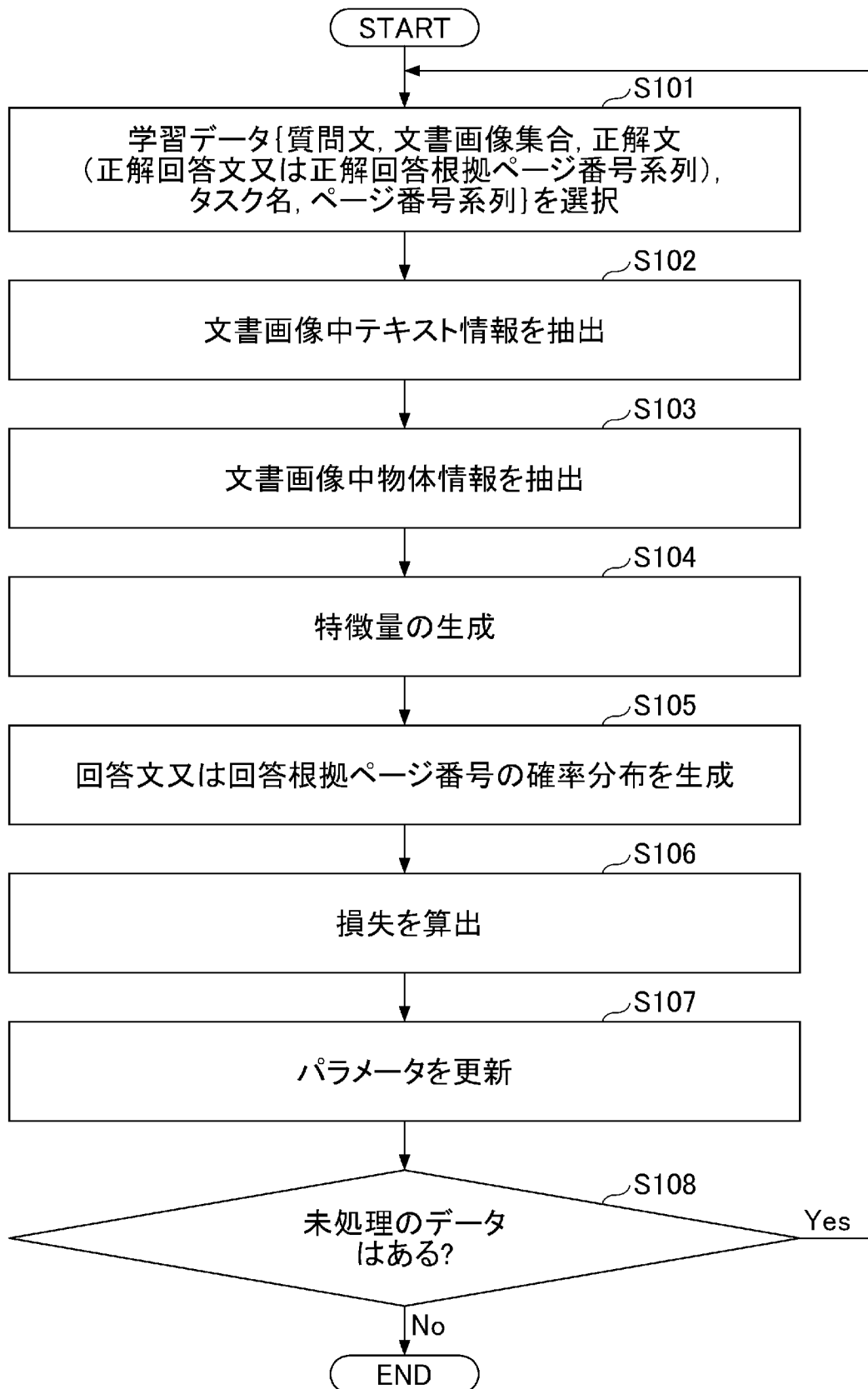


p.5

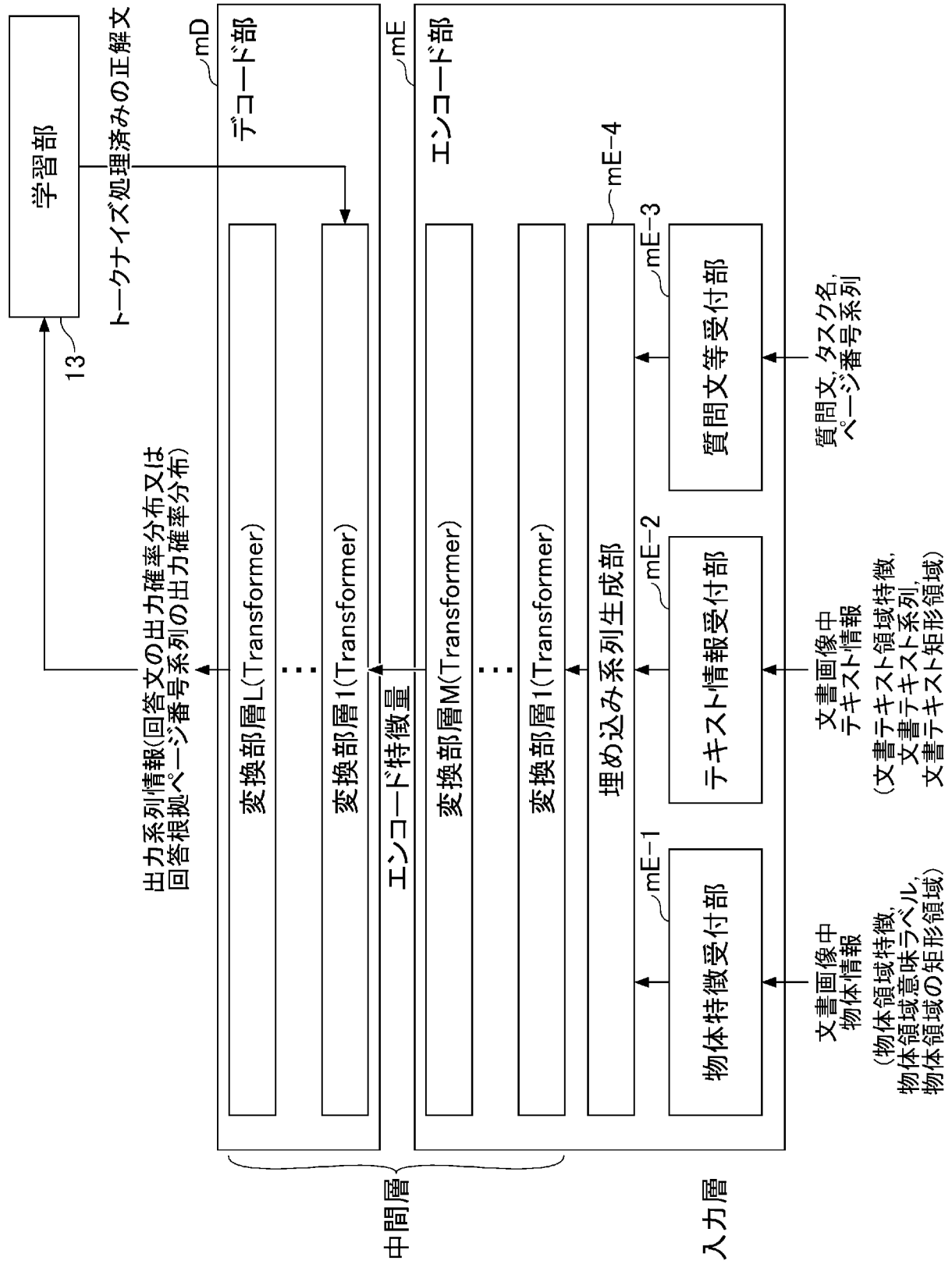


タスク名=「Evidence Selection」の場合の正解文(=正解回答根拠ページ番号系列): 4, 5

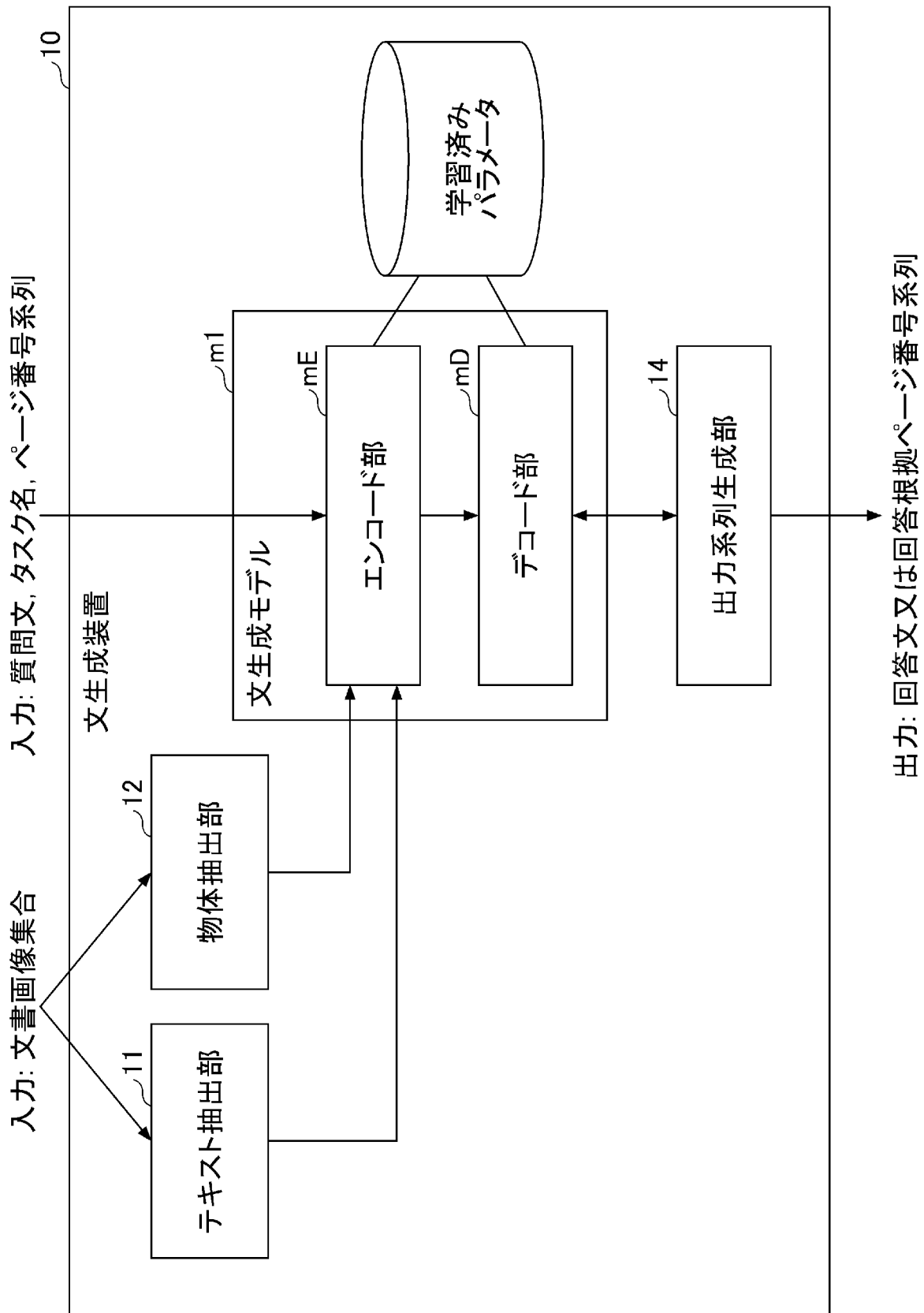
[図4]



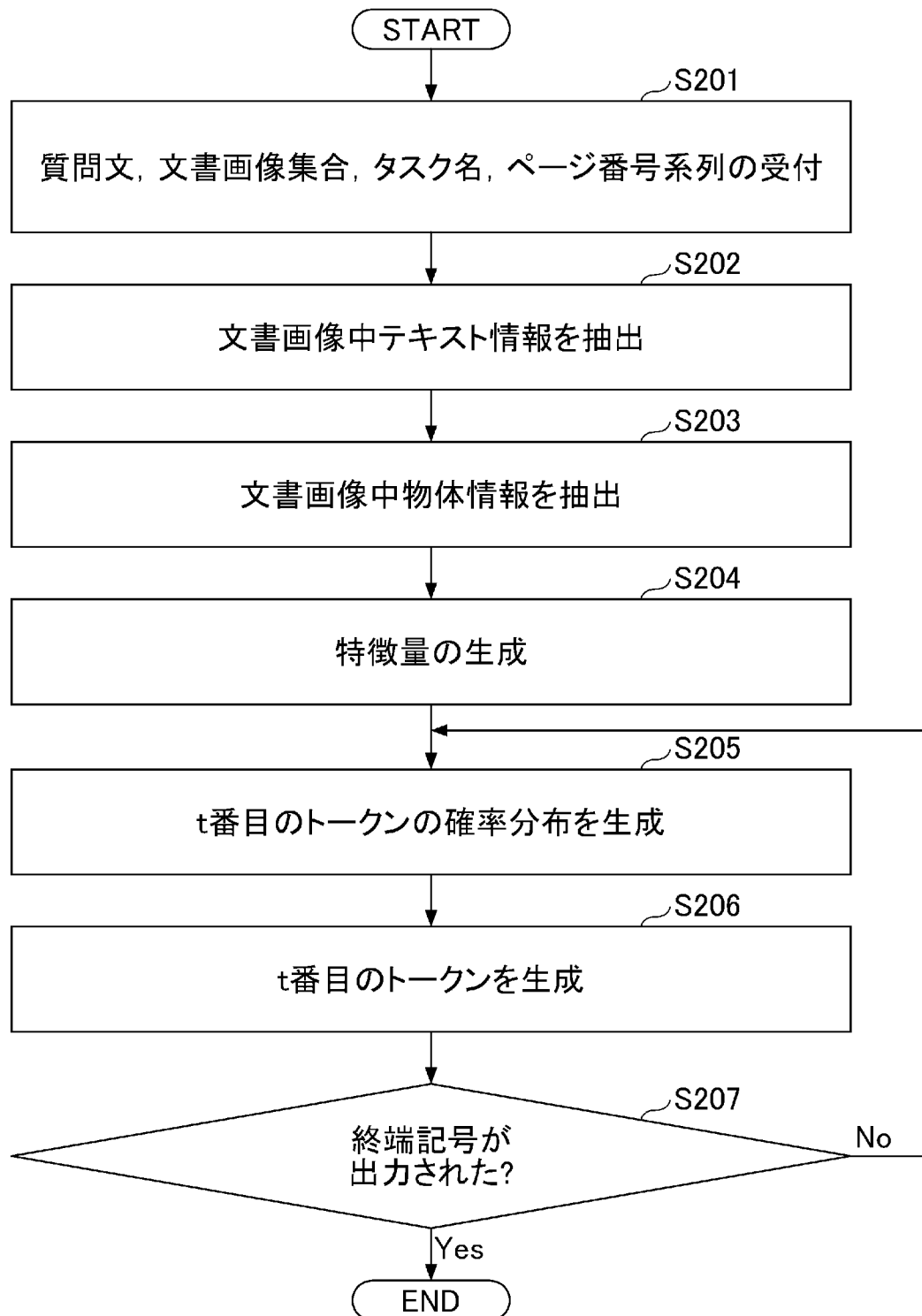
[図5]



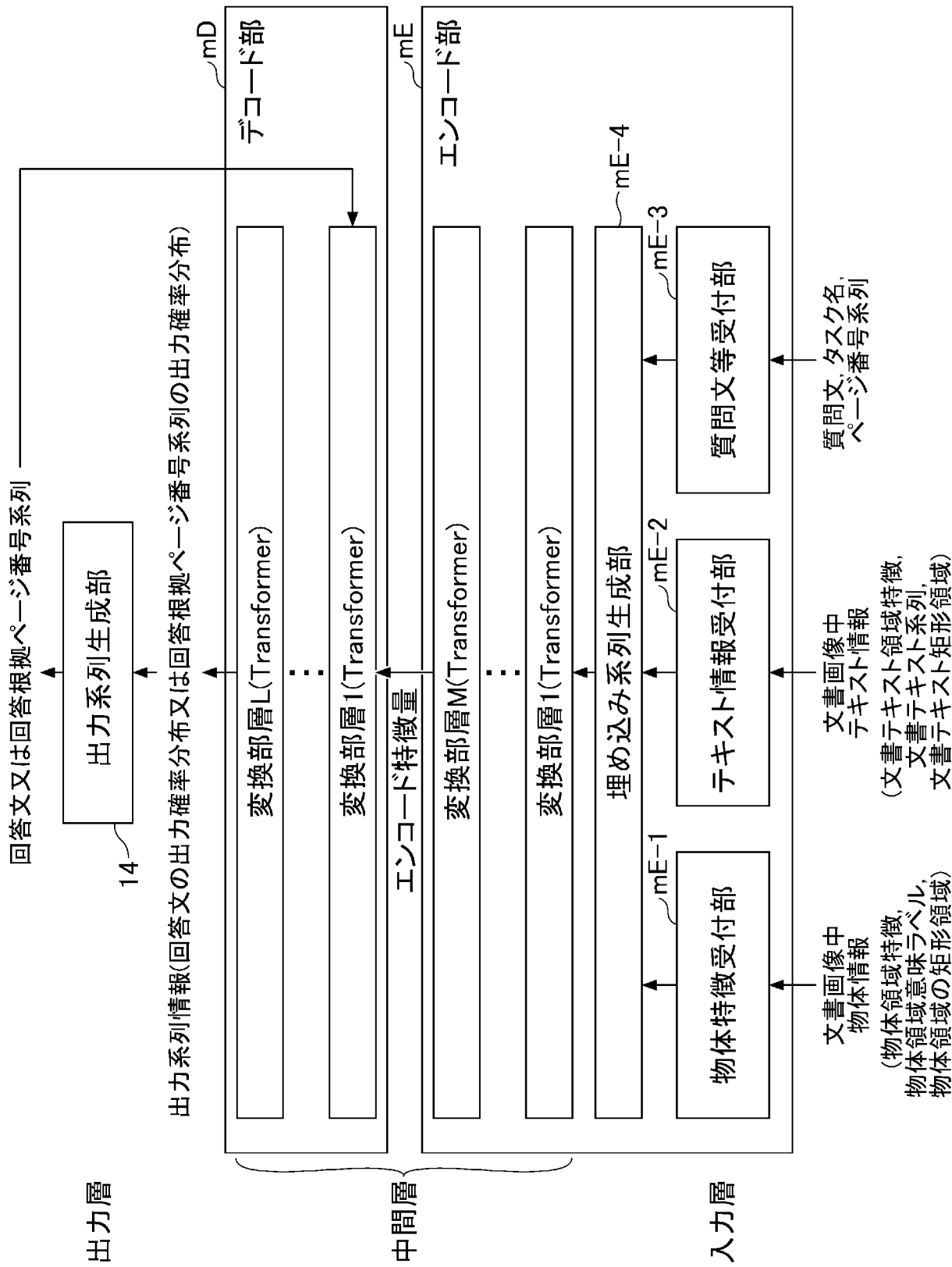
[図6]



[図7]



[図8]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2022/047446

A. CLASSIFICATION OF SUBJECT MATTER**G06N 20/00**(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)

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 1971-2023

Registered utility model specifications of Japan 1996-2023

Published registered utility model applications of Japan 1994-2023

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.
Y	TITO, R. et al., Hierarchical multimodal transformers for Multi-Page DocVQA, arXiv:2212.05935v 1 [cs.CV], 07 December 2022, pp. 1-14, [online], [retrieved on 07 March 2023], Retrieved from <https://arxiv.org/abs/2212.05935> <doi: 10.48550/ARXIV.2212.05935>	1, 3-7
A	pp. 1-14	2
Y	GUO, Q. et al., Key-Region and Layout Learning for Contract Intelligent Identification, 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), 08 February 2022, pp. 57-61, [online], [retrieved on 07 March 2023], Retrieved from <https://ieeexplore.ieee.org/document/9696614> <doi: 10.1109/ICESIT.2021.9696614>	1, 3-7
A	pp. 57-61	2

☒ 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

10 March 2023

Date of mailing of the international search report

20 March 2023

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

International application No.

PCT/JP2022/047446

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	PRAMANIK, S. et al., Towards a Multi-modal, Multi-task Learning based Pre-training Framework for Document Representation Learning, arXiv:2009.14457v 2 [cs.CL], 05 January 2022, pp. 1-11, [online], [retrieved on 07 March 2023], Retrieved from <https://arxiv.org/abs/2009.14457> <doi: 10.48550/ARXIV.2009.14457> pp. 1-11	1-7

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 1971-2023年 日本国実用新案登録公報 1996-2023年 日本国登録実用新案公報 1994-2023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
Y	TITO, R. et al., Hierarchical multimodal transformers for Multi-Page DocVQA, arXiv:2212.05935v1 [cs.CV], 2022.12.07, pp. 1-14, [online], [retrieved on 2023-03-07], Retrieved from <https://arxiv.org/abs/2212.05935> <doi: 10.48550/ARXIV.2212.05935>	1, 3-7
A	pp. 1-14	2
Y	GUO, Q. et al., Key-Region and Layout Learning for Contract Intelligent Identification, 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), 2022.02.08, pp. 57-61, [online], [retrieved on 2023-03-07], Retrieved from <https://ieeexplore.ieee.org/document/9696614> <doi: 10.1109/ICESIT53460.2021.9696614>	1, 3-7
A	pp. 57-61	2
☒ 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、一般的技术水準を示すもの “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 10.03.2023		国際調査報告の発送日 20.03.2023
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） 打出 義尚 5B 4440 電話番号 03-3581-1101 内線 3546

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	PRAMANIK, S. et al., Towards a Multi-modal, Multi-task Learning based Pre-training Framework for Document Representation Learning, arXiv:2009.14457v2 [cs.CL], 2022.01.05, pp. 1-11, [online], [retrieved on 2023-03-07], Retrieved from <https://arxiv.org/abs/2009.14457> <doi: 10.48550/ARXIV.2009.14457> pp. 1-11	1-7