

(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
2021年11月25日(25.11.2021)



(10) 国際公開番号

WO 2021/234950 A1

(51) 国際特許分類:

G06F 40/253 (2020.01)

東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP).

(21) 国際出願番号:

PCT/JP2020/020334

(74) 代理人: 中尾 直樹, 外 (NAKAO, Naoki et al.);

(22) 国際出願日:

2020年5月22日(22.05.2020)

〒1600022 東京都新宿区新宿三丁目1番22号 新宿NSOビル6階 Tokyo (JP).

(25) 国際出願の言語:

日本語

(26) 国際公開の言語:

日本語

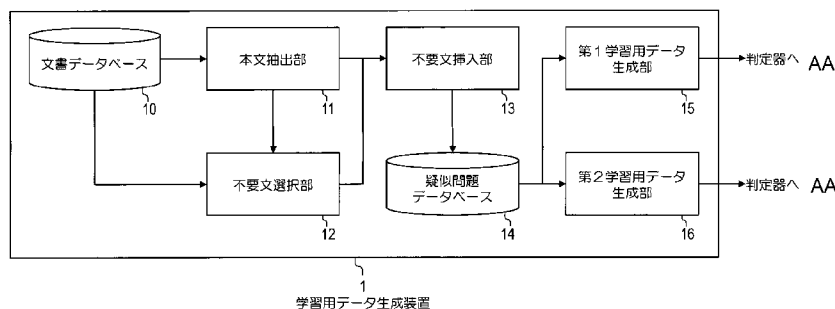
(71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(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, 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.

(72) 発明者: 成松 宏美 (NARIMATSU, Hiromi); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP). 杉山 弘晃 (SUGIYAMA, Hiroaki); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP). 東中 竜一郎 (HIGASHINAKA, Ryuichiro); 〒1808585

(54) Title: LEARNING DATA GENERATION DEVICE, ASSESSMENT DEVICE, LEARNING DATA GENERATION METHOD, ASSESSMENT METHOD, AND PROGRAM

(54) 発明の名称: 学習用データ生成装置、判定装置、学習用データ生成方法、判定方法、プログラム



1 ... Learning data generation device
10... Document database
11... Main text extraction unit
12... Unnecessary phrase selection unit
13... Unnecessary phrase insertion unit
14... Pseudo problem database
15... First learning data generation unit
16... Second learning data generation unit
AA ... To assessment unit

(57) Abstract: Provided is a learning data generation device that is able to generate learning data used in learning of an assessment unit for assessing which portion of a document is unnatural. The learning data generation device includes: a main text extraction unit for extracting a portion of a document as a main text; an unnecessary phrase selection unit for selecting, as an unnecessary phrase, one phrase in the document that is not included in the main text; an unnecessary phrase insertion unit for generating a pseudo problem by inserting the unnecessary phrase into the main text; a first learning data generation unit for generating first learning data by setting the unnecessary phrase as an option, and attaching to the pseudo problem a first label signifying that the context of the passage encompassing the option is natural; and a second learning data generation unit for generating second learning data by setting one phrase from the main text as an option, and attaching to the pseudo problem a second label signifying that the context of the passage encompassing the option is unnatural.

(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))

(57) 要約: 文書のどの部分が不自然であるかを判定するための判定器の学習に用いる学習用データを生成することができる学習用データ生成装置を提供する。文書の一部を本文として抽出する本文抽出部と、文書のうち本文に含まれない一文を不要文として選択する不要文選択部と、本文に不要文を挿入して疑似問題を生成する不要文挿入部と、不要文を選択肢とし、選択肢の前後に存在する文章のつながりが自然であることを意味する第1のラベルを疑似問題に付して、第1の学習用データを生成する第1学習用データ生成部と、本文のうち一文を選択肢とし、選択肢の前後に存在する文章のつながりが不自然であることを意味する第2のラベルを疑似問題に付して、第2の学習用データを生成する第2学習用データ生成部を含む。

明 細 書

発明の名称：

学習用データ生成装置、判定装置、学習用データ生成方法、判定方法、プログラム

技術分野

[0001] 本発明は、文章の流れの自然さを判定するための学習用データ生成装置、判定装置、学習用データ生成方法、判定方法、プログラムに関する。

背景技術

[0002] 自然言語処理モデルの従来技術として、例えばBERT (Bidirectional Encoder Representations from Transformers、非特許文献1) による双方向のエンコード表現やXLNet (非特許文献2) などがある。

先行技術文献

非特許文献

- [0003] 非特許文献1：“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”，Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Google AI Language, arXiv 24 May 2019
- 非特許文献2：“XLNet: Generalized Autoregressive Pretraining for Language Understanding”，Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, arXiv 2 Jan 2020

発明の概要

発明が解決しようとする課題

- [0004] 従来の言語モデルは、自然な文書を教師データとして用いて予め学習することで、文書全体の自然な文書らしさのスコアを算出することができる。しかしながら従来の言語モデルは、文書全体の評価はできても、文書のどの部分が不自然であるかを判定することはできない。
- [0005] そこで本発明では、文書のどの部分が不自然であるかを判定するための判

定器の学習に用いる学習用データを生成することができる学習用データ生成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

[0006] 本発明の学習用データ生成装置は、本文抽出部と、不要文選択部と、不要文挿入部と、第1学習用データ生成部と、第2学習用データ生成部を含む。

[0007] 本文抽出部は、文書の一部を本文として抽出する。不要文選択部は、文書のうち本文に含まれない一文を不要文として選択する。不要文挿入部は、本文に不要文を挿入して疑似問題を生成する。第1学習用データ生成部は、不要文を選択肢とし、選択肢の前後に存在する文章のつながりが自然であることを意味する第1のラベルを疑似問題に付して、第1の学習用データを生成する。第2学習用データ生成部は、本文のうちの一文を選択肢とし、選択肢の前後に存在する文章のつながりが不自然であることを意味する第2のラベルを疑似問題に付して、第2の学習用データを生成する。

発明の効果

[0008] 本発明の学習用データ生成装置によれば、文書のどの部分が不自然であるかを判定するための判定器の学習に用いる学習用データを生成することができる。

図面の簡単な説明

[0009] [図1]文書、本文、不要文、疑似問題の例1を示す図。

[図2]文書、本文、不要文、疑似問題の例2を示す図。

[図3]疑似問題をBERT/XLNet入力フォーマットに変換する例1を示す図。

[図4]疑似問題をBERT/XLNet入力フォーマットに変換する例2を示す図。

[図5]実施例1の学習用データ生成装置の構成を示すブロック図。

[図6]実施例1の学習用データ生成装置の疑似問題生成動作を示すフローチャート。

[図7]実施例1の不要文選択部の構成を示すブロック図。

[図8]実施例1の不要文選択部の動作を示すフローチャート。

[図9]実施例1の不要文挿入部の構成を示すブロック図。

[図10]実施例1の不要文挿入部の動作を示すフローチャート。

[図11]実施例1の学習用データ生成装置の学習用データ生成動作を示す図。

[図12]実施例1の判定装置の構成を示すブロック図。

[図13]実施例1の判定装置の動作例1を示すフローチャート。

[図14]実施例1の判定装置の動作例2を示すフローチャート。

[図15]疑似問題をBERTに入力して出力Class Labelを得る例を示す図。

[図16]第1、第2の入力フォーマットを例示する図。

[図17]第3の入力フォーマットを例示する図。

[図18]第4の入力フォーマットを例示する図。

[図19]第5の入力フォーマットを例示する図。

[図20]コンピュータの機能構成例を示す図。

発明を実施するための形態

[0010] 以下、本発明の実施の形態について、詳細に説明する。なお、同じ機能を有する構成部には同じ番号を付し、重複説明を省略する。

実施例 1

[0011] 以下、図1から図4を参照して、実施例1における疑似問題の生成および入力フォーマットの作成について説明する。

[0012] <処理1>

まず、つながりが自然であるか否かを判断したい文（以下、本文という）を含み、本文よりも少し長い程度の文（以下、文書という、文書は例えば一つの短文、一つの段落程度の長さとする）から、N個（Nは1以上の整数）の文（以下、N文ともいう）を抽出して本文とする。文書の例を図1の左側の枠内の文書として、本文の例を太字で例示する。

[0013] <処理2>

次に、文書の本文以外の箇所から不要な文（以下、不要文という）として挿入する文を選択する。選択はランダムに行われてもよいし、他の基準で行われてもよい（詳細は後述）。不要文の例を同図に下線で示す。

[0014] <処理3>

次に、不要文を本文中の何れかの文と文の間（文間）に挿入する。挿入位置はランダムに決定されてもよいし、他の基準で決定されてもよい（詳細は後述）。このように生成した複数の文を疑似問題と呼び、疑似問題と本文、不要文、不要文の挿入位置を紐づけて、ラベルを付して学習用データとして保持する。

[0015] 図2は、図1と同じ文書内（「かぐや姫」）で別の疑似問題を生成した例を示したものである。同図に示すように、文書からすでに作成済の疑似問題の本文に合致しない連続するN文を選択して本文として抽出し、処理2～3を繰り返す。

[0016] <BERT/XLNet入力フォーマット>

図3、図4は、疑似問題をBERT/XLNet入力フォーマットに変換した例を示す図である。入力フォーマットというのは、図3、図4に示すtokensの形式のことである。ここで、CLSは文頭、SEPはセパレータを示すIDである。

[0017] [学習用データ生成装置1]

以下、図5を参照して、実施例1の学習用データ生成装置1の構成を説明する。同図に示すように本実施例の学習用データ生成装置1は、文書データベース10と、本文抽出部11と、不要文選択部12と、不要文挿入部13と、疑似問題データベース14と、第1学習用データ生成部15と、第2学習用データ生成部16を含む。

[0018] 以下、図6を参照して、本実施例の学習用データ生成装置1の疑似問題生成動作を説明する。

[0019] <文書データベース10>

文書データベース10には、あらかじめ一般的な文書が蓄積されている。蓄積される文書は例えばWEB上から取得することができる。文書の長さが指定する範囲に含まれる場合、同図に示すステップS11～S14が実行される。

[0020] <本文抽出部11>

本文抽出部11は、文書データベース10に蓄積された文書のうちのいずれ

れかの文書を入力とし、文書の一部であって、疑似問題に未使用の連続するN文を本文として抽出し、本文を出力する（S 1 1）。

[0021] ＜不要文選択部 1 2＞

不要文選択部 1 2 は、ステップ S 1 1 の処理を実行した文書と、ステップ S 1 1 で抽出された本文を入力とし、文書のうち本文に含まれない一文を不要文として選択し、不要文を出力する（S 1 2）。なお、不要文選択部 1 2 は、不要文をランダムに選択してもよいし、文書のうちの本文に含まれない一文のうち、本文との文間類似度が高い一文を不要文として選択してもよい（後述するステップ S 1 2 1 ～ S 1 2 5 を参照）。

[0022] ＜不要文挿入部 1 3＞

不要文挿入部 1 3 は、ステップ S 1 1 で出力された本文と、ステップ S 1 2 で出力された不要文を入力とし、本文の何れかの文間に不要文を挿入して疑似問題を生成し、疑似問題と、不要文の挿入位置を出力する（S 1 3）。なお、不要文挿入部 1 3 は、不要文の挿入位置をランダムに決定してもよいし、本文のうちの隣接する文のペアのうち、不要文との文間類似度が高いペアの間に不要文を挿入してもよい（後述するステップ S 1 3 1 ～ S 1 3 5 を参照）。

[0023] ＜疑似問題データベース 1 4＞

疑似問題データベース 1 4 は、ステップ S 1 3 で出力された疑似問題と、不要文の挿入位置とを入力とし、疑似問題と、不要文の挿入位置とを紐づけて蓄積する（S 1 4）。

[0024] ＜繰り返し処理＞

ステップ S 1 1 で指定された文書に本文として使用していない連続するN文が存在する場合、ステップ S 1 1 ～ S 1 4 が繰り返し実行される。

[0025] ＜不要文選択部 1 2 の詳細な構成例と動作＞

以下、図 7 を参照して、不要文選択部 1 2 の詳細な構成例を示す。同図に示すように、本実施例の不要文選択部 1 2 は、第 1 文ベクトル算出部 1 2 1 と、第 2 文ベクトル算出部 1 2 2 と、文間類似度算出部 1 2 3 と、一時保持

部 1 2 4 と、選択部 1 2 5 を含む。

[0026] 以下、図 8 を参照して、本実施例の不要文選択部 1 2 の動作を説明する。

[0027] ≪第 1 文ベクトル算出部 1 2 1≫

第 1 文ベクトル算出部 1 2 1 は、本文を入力とし、本文の文ベクトルを算出し、本文の文ベクトルを出力する（S 1 2 1）。

[0028] ≪第 2 文ベクトル算出部 1 2 2≫

第 2 文ベクトル算出部 1 2 2 は、文書と本文を入力とし、文書中の本文以外の、まだ文ベクトルを算出していない文を一文選択し、文ベクトルを算出し、文ベクトルを出力する（S 1 2 2）。

[0029] ≪文間類似度算出部 1 2 3≫

文間類似度算出部 1 2 3 は、ステップ S 1 2 1 で算出された本文の文ベクトルと、ステップ S 1 2 2 で算出された一文の文ベクトルを入力とし、本文の文ベクトルと一文の文ベクトルから文間類似度を算出し、文間類似度を出力する（S 1 2 3）。

[0030] ≪一時保持部 1 2 4≫

一時保持部 1 2 4 は、ステップ S 1 2 2 で文ベクトルの算出対象となった一文と、対応する文間類似度を入力とし、ステップ S 1 2 2 で文ベクトルの算出対象となった一文が最初の文である場合、または、ステップ S 1 2 2 で文ベクトルの算出対象となった一文の本文との文間類似度が、これまでの一文の本文との文間類似度よりも高い場合、対象の一文を本文との文間類似度が最大のものとして保持する（S 1 2 4）。

[0031] 本文以外の文書中にまだ文ベクトルを算出していない一文がある場合には、ステップ S 1 2 2 に戻る。本文以外の文書中にまだ文ベクトルを算出していない一文が存在しない場合には、一時保持部 1 2 4 は、文間類似度が最大のものとして保持している一文を出力する。

[0032] ≪選択部 1 2 5≫

選択部 1 2 5 は、文間類似度が最大のものとして保持されている一文を入力とし、これを不要文として選択して、出力する（S 1 2 5）。

[0033] <不要文挿入部 13 の詳細な構成例と動作>

以下、図 9 を参照して、不要文挿入部 13 の詳細な構成例を示す。同図に示すように、本実施例の不要文挿入部 13 は、不要文ベクトル算出部 131 と、隣接文ベクトル算出部 132 と、文間類似度算出部 133 と、一時保持部 134 と、選択部 135 を含む。

[0034] 以下、図 10 を参照して、本実施例の不要文挿入部 13 の動作を説明する。

[0035] ≪不要文ベクトル算出部 131≫

不要文ベクトル算出部 131 は、不要文を入力とし、不要文の文ベクトルを算出し、不要文の文ベクトルを出力する (S 131)。

[0036] ≪隣接文ベクトル算出部 132≫

隣接文ベクトル算出部 132 は、本文を入力とし、本文中の、まだ文ベクトルを算出していない隣接する文のペアを選択し、隣接する文の各文ベクトルを算出し、隣接する文の各文ベクトルを出力する (S 132)。

[0037] 本文中にまだ文ベクトルを算出していない隣接する文のペアがある場合には、隣接文ベクトル算出部 132 は、ステップ S 132 を繰り返し実行する。本文中にまだ文ベクトルを算出していない隣接する文のペアがない場合には、ステップ S 133 に進む。

[0038] ≪文間類似度算出部 133≫

文間類似度算出部 133 は、ステップ S 131 で算出された不要文の文ベクトルと、ステップ S 132 で算出された隣接する文の各文ベクトルを入力とし、不要文の文ベクトルと隣接する文の各文ベクトルから文間類似度を算出し、文間類似度を出力する (S 133)。

ステップ S 133 の詳細な処理方法を以下に示す。不要文の文ベクトルをベクトル 1 とした場合、2 つの方法が考えられる。

方法 1) 隣接する文ペア (文 1, 文 2) の文ベクトルを足し合わせた (もしくは平均値をとった) 値をベクトル 2 とする。文間類似度算出部 133 は、ベクトル 1 とベクトル 2 の文間類似度 (たとえばコサイン類似度) を算出し

て出力する。

方法2) 隣接する文ペア(文1, 文2)の各文ベクトルを算出し、ベクトル4、ベクトル5とする。文間類似度算出部133は、ベクトル1とベクトル4、ベクトル1とベクトル5の文間類似度をそれぞれ算出し、その平均値を文間類似度として出力する。

[0039] <<一時保持部134>>

一時保持部134は、ステップS133で算出された文間類似度と、これに対応する隣接する文を入力とし、ステップS133が一回目の処理である場合、または、ステップS133で算出された文間類似度が、これまでの文間類似度よりも高い場合、対象の隣接する文の間の地点を文間類似度が最大の地点として保持する(S134)。

[0040] 本文中にまだ文間類似度を算出していない隣接する文のペアがある場合には、ステップS133に戻る。本文中にまだ文間類似度を算出していない隣接する文のペアが存在しない場合には、一時保持部134は、文間類似度が最大のものとして保持している地点を出力する。

[0041] <<選択部135>>

選択部135は、本文と、不要文と、文間類似度が最大のものとして保持されている地点を入力とし、当該地点を不要文の挿入位置として選択し、本文に不要文を挿入して、疑似問題を生成し、出力する(S135)。

図10は、ステップS132を最初に繰り返すことで、本文中の全ての隣接する文のペアについて文ベクトルをまとめて算出する手法であるが、これに限定されず、ステップS132において、一つの隣接する文のペアに対して文ベクトルを算出した後、ステップS133において対応する文間類似度を算出し、算出した文間類似度が保持している文間類似度よりも高い場合、あるいは、文間類似度の算出が初回である場合に、ステップS134の保持を行い、次の隣接する文のペアに対して文ベクトルを算出して、対応する文間類似度を算出し、算出した文間類似度が保持している文間類似度よりも高い場合に、ステップS134の保持を行い、…という処理を繰り返し実行し

てもよい。

[0042] 以下、図 1 1 を参照して、本実施例の学習用データ生成装置 1 の学習用データ生成動作を説明する。

[0043] <第 1 学習用データ生成部 1 5>

第 1 学習用データ生成部 1 5 は、疑似問題データベース 1 4 に蓄積された疑似問題を入力とし、この疑似問題の不要文を選択肢とし、選択肢の前後に存在する文章のつながりが自然であることを意味する第 1 のラベルを当該疑似問題に付して、第 1 の学習用データを生成し、第 1 の学習用データを出力する（S 1 5）。

[0044] <第 2 学習用データ生成部 1 6>

第 2 学習用データ生成部 1 6 は、疑似問題データベース 1 4 に蓄積された疑似問題を入力とし、この疑似問題の本文のうちの一文を選択肢とし、選択肢の前後に存在する文章のつながりが不自然であることを意味する第 2 のラベルを当該疑似問題に付して、第 2 の学習用データを生成し、第 2 の学習用データを出力する（S 1 6）。

[0045] 学習していない疑似問題と選択肢のペアがある場合には、学習していない疑似問題と選択肢のペアに対してステップ S 1 5、S 1 6 が再度実行される。

[0046] <学習用データ生成部で用いる入力フォーマット>

上記の第 1 学習用データ生成部 1 5 と第 2 学習用データ生成部 1 6 で使用された入力フォーマット（選択肢となる一文を選び出し、選択肢の前後の文を連結）は、あくまで一例である。上記の例に限定されず、学習用データ生成装置 1 の学習用データ生成部は、所定の入力フォーマットに従った文章と、文章のつながりが自然であるか不自然であるかのラベルを含むように、学習用データを生成すればよい。所定の入力フォーマットの例（入力フォーマット（1）～（5））については後述する。

[0047] [判定装置 2]

以下、図 1 2 を参照して、実施例 1 の判定装置 2 の構成を説明する。同図

に示すように本実施例の判定装置 2 は、入力フォーマット作成部 2 1 と、判定器 2 1 A と、尤度取得部 2 2 と、判定部 2 3 を含む。

[0048] 以下、図 1 3、図 1 4 を参照して、本実施例の判定装置 2 の動作を説明する。

[0049] <入力フォーマット作成部 2 1>

入力フォーマット作成部 2 1 は、判定対象の文章を入力とし、判定対象の文章から所定の入力フォーマットに従った判定用データを作成して、当該判定用データを出力する（S 2 1）。所定の入力フォーマットのバリエーション（入力フォーマット（1）～（5））については後述する。

[0050] <判定器 2 1 A>

判定器 2 1 A は、所定の入力フォーマットに従った文章と、文章のつながりが自然であるか不自然であるかのラベルを含む学習用データを用いて予め学習されているものとし、ステップ S 2 1 で作成した判定用データを入力とし、判定対象の文章のつながりが自然であるか否かの尤度を出力する（S 2 1 A）。

[0051] <尤度取得部 2 2>

尤度取得部 2 2 は、判定器 2 1 A から判定対象の文章のつながりが自然であるか否かの尤度を取得して、当該尤度を出力する（S 2 2）。

[0052] <判定部 2 3>

判定部 2 3 は、尤度を入力として、尤度に基づいて判定対象の文章のつながりが自然であるか否かを判定し、判定結果を出力する（S 2 3）。ステップ S 2 3 に関しては、図 1 3 に示す動作例 1 のように、尤度とあらかじめ定めた閾値を比較して、尤度が閾値以上とならない場合に、判定対象の文章を自然な文でないと判定し（S 2 3-1）、尤度が閾値以上となる場合に、判定対象の文章を自然な文であると判定してもよい（S 2 3-2）。

[0053] またステップ S 2 3 に関しては、図 1 4 に示す動作例 2 のように、自然かどうかの尤度がこれまでで最も高かった場合に、対象の選択肢と尤度を保持し（S 2 3-3）、尤度を算出していない選択肢がなくなった場合に、現在

保持している選択肢に対応する文章を自然な文と判定してもよい（S 23-4）。このとき、自然かどうかの尤度は判定結果の信頼度として用いることができる。

[0054] <BERT>

以下、図 15 を参照してBERTについて説明する。BERTには、大量の自然文書を学習させたPretrainedモデル（汎用言語モデル）と、2文の結束性や含意関係、系列ラベリングなどを判定できるように適応（転移学習）させたfinetuneモデルの2種類が存在する。なお、BERTの出力は、クラスと尤度である。

[0055] BERTを用いるメリットは、大量の自然文書で言語モデルを学習させたPretrainedモデルを用いることで、解きたいタスクに対してそこまで大量の学習データがなくても転移学習によりある程度の精度の正解率を達成できることである。

[0056] BERTにおいて、2文の結束性（つながりが自然かどうか）を判定する判定器は同図に示す構造を持つ。

[0057] 学習および判定時の入出力および手順1)～5)について以下に説明する。

[0058] 1) BERTの中で結束性を判定するように転移学習するための入出力形式を用いる。

[0059] 2) 同図に示すように文1と文2の2文を用意し、セパレータ[SEP]で区切って入力する。入力の最初の[CLS]はクラスを表すベクトルである。例えば0, 1で判定する場合には、0, 1およびどちらのクラスが入力データのクラスとして正しいかの情報を入力する。

[0060] 3) クラスラベルは例えば、0（自然）、1（不自然）のように割り当てる。例えば、文1と文2が文章中の連続する2文であるとき、CLSラベルは0（自然）、ランダムに選択した2文であるときは、CLSラベルは1（不自然）とする。

[0061] 4) クラスラベルおよびその尤度を出力する。

[0062] 5) 判定時はCLSラベルに正解の入力を含めない以外は、同様の入出力となる。

[0063] 同図の[CLS]はクラスラベルであり、入力の先頭に配置する。[SEP]は、入力の区切り位置であることを示す。SegmentIDには、0か1が入る。一般に1を立てたところは、0の箇所と異なる性質をもつものであることを表す（ことが期待されて与えられたIDである）。例えば含意関係を判定するとき、文1に対して0を、文2に対して1を与える。

[0064] EはEmbeddingされた記号で、Tokenを数字で表されるIDに変換したものであり、SentencePieceなどを用いて変換する。

[0065] <本実施例におけるBERTの転移学習方法>

引き続き図15を参照して、本実施例におけるBERTの応用方法を説明する。BERTを用いて、文全体のつながりが自然かどうかを判定するようにデータおよび判定器を作成する場合、<BERT>で説明した、BERTの結束性判定の転移学習を行う入力形式を用いて、今回あらたに作成したデータを学習させる。

[0066] 学習手順は以下のi)～iii)のとおりである。

[0067] i) 例えば、同図左に示す疑似問題を用いて学習する場合、0（自然）のラベルが付与されるデータと1（不自然）のラベルが付与されるデータの2種類を作成できる。

[0068] ii) 0（自然）のラベルが付与されるデータの作成方法：不要文を選択肢として、選択肢より前のN文（図の本文の場合にはN＝1～3）を文1、選択肢より後ろのN文（図の本文の場合にはN＝3～6）を文2として入力する。

[0069] iii) 1（不自然）のラベルが付与されるデータの作成方法：不要文以外の文を選択肢（例えばもともとある文4）として、選択肢より前のN文（図の本文の場合にはN＝1～挿入した文まで）を文1、選択肢より後ろのN文（図の本文の場合にはN＝5～6）を文2として入力する。

[0070] ただし、同図に示す入力フォーマットは一例であり、他の入力フォーマッ

トも使用可能である。

[0071] BERTなどを用いた入力形式について、2つの着想が考えられる。一つの着想として、不要なものを見つけるという考え方（対象が不要かどうか、含意関係があるかを判定するという考え方）がある。もう一つの着想として、文書の流れの自然さを判定するという考え方（2つの文の繋がり goodness を判定するという考え方）がある。

[0072] 不要なものを見つける方法として、判定時の工夫のみを検討してもよい。すなわち、学習を行わずに言語モデルの尤度だけを用いて、不要なものを見つけるという発想である。

[0073] <入力フォーマットの使い方>

1つの方法は、入力フォーマットごとに異なる判定器を作成する方法である。

[0074] もう一つの方法は、複数のモデルを組み合わせアンサンブル学習をする方法である。「アンサンブル学習」とは、入力フォーマットごとに異なる判定器を作成しておき、入力フォーマットを判定器に合わせた入力を用いて判定し、各判定器の出力する尤度を重みづけ加算するなど統合するという意味である。

[0075] その場合には、各入力フォーマットで尤度を算出し、それぞれに重み付けをして算出したスコア（統合した尤度）を用いて、学習および判定を行う。

[0076] 以降の入力形式では、BERTへの入力を想定しSegmentIDも含めて説明するが、BERTに限らず、XLNet等も使用できる。BERTの進化系として提案されたXLNetなどではSegmentIDを用いないため、SegmentIDはなくても良いものとする。

[0077] 以下、入力フォーマットの例を図16～図19を参照して説明する。

[0078] <入力フォーマット（1）>

入力フォーマット（1）は、図16に示すように、疑似問題のうちの一文を選択肢として、選択肢を除く疑似問題全体と選択肢とを連結して生成する。このフォーマットを用いることにより、文書全体と選択肢の類似性や含意関係が学習され、無関係な文に対して、不要を判断できる。

[0079] ＜入力フォーマット（２）＞

入力フォーマット（２）は、図１６に示すように、疑似問題の各文を各選択肢とし、各選択肢から選びだした一つを選択肢と、選び出されなかったすべての選択肢からなる選択肢群とを連結して生成する。このフォーマットを用いることにより、文書全体ではなく、選択肢のみに限定し比較することで、選択肢間の違いを分類する学習が効率的に行える。

[0080] ＜入力フォーマット（３）＞

入力フォーマット（３）は、図１７に示すように、疑似問題のうちの一文を選択肢として、疑似問題のうちの選択肢よりも前に存在する文と、選択肢と、疑似問題のうちの選択肢よりも後に存在する文とを連結して生成する。このフォーマットを用いることにより、選択肢前後のつながりの自然さを評価できる。

[0081] ＜入力フォーマット（４）＞

入力フォーマット（４）は、図１８に示すように、疑似問題のうちの一文を選択肢として、疑似問題のうちの選択肢よりも前に存在する文と、疑似問題のうちの選択肢よりも後に存在する文とを連結して生成する。入力フォーマット（４）－１は、選択肢よりも後に存在する文にSegmentID=1を付すことにより、選択肢を抜いたときの後ろにつながる文としての妥当性を表現できる。入力フォーマット（４）－２は、全体にSegmentID=0を付すことにより、選択肢を抜いたときの全体のつながりの自然さを表現できる。

[0082] ＜入力フォーマット（５）＞

入力フォーマット（５）は、図１９に示すように、疑似問題のうちの一文を選択肢として、疑似問題のうちの選択肢よりも前に存在するN個（Nは1以上の整数）の文と、選択肢および選択肢よりも後に存在するN個の文を含む文とを連結した入力フォーマットAと、疑似問題のうちの選択肢よりも前に存在するN個（Nは1以上の整数）の文と、選択肢よりも後に存在するN+1個の文を含む文とを連結した入力フォーマットBの双方からなる。

[0083] 入力フォーマットAから取得された尤度を S_a とし、入力フォーマットBから

取得された尤度を S_b とすれば、選択肢＝不自然な文である場合に、尤度 S_a と尤度 S_b の尤度差 $|S_a - S_b|$ が大きくなるため、尤度差 $|S_a - S_b|$ が大きい選択肢を不自然な文と判断することができる。

[0084] 実施例に記載の装置、方法は、不要な一文を選択するだけでなく、正しい文がどれかを選ぶ選択式の問題や、選択肢が無いような実世界の文章（例えば、エッセイなどにおいて綺麗な文書をかけているかを判定するようなタスク）にも応用できる。

[0085] その場合には、判定対象の文章と注目している一文および挿入位置のペアは以下ようになる。

[0086] （１）不要文を判定する場合：

（判定対象の文章、不要かどうか判定する一文、不要かどうか判定する一文の位置）を生成し、自然かどうかの尤度を算出する。尤度が閾値以下であれば、不要な文と判断する。択一式の問題であれば、尤度の算出を各文に対して実施し、最も尤度が低い文を不要な文であると判定する。

[0087] （２）必要な文を判定する場合：

（判定対象の文章、必要かどうか判定する一文、必要かどうか判定する一文の位置）を生成し、自然かどうかの尤度を算出する。尤度が閾値以上であれば、必要な文と判断する。択一式の問題であれば、尤度の算出を各文に対して実施し、最も尤度が高い文を必要な文であると判定する。

[0088] （３）判定対象の文章が綺麗かを判定する場合：

（判定対象の文章、判定対象の文章から選択した任意の一文、判定対象の文章から選択した任意の一文の位置）を判定対象の文章中の任意の文全てについて生成し、それぞれの自然かどうかの尤度を算出する。尤度が閾値以上であれば、判定対象の文章が自然な文章であると判定する。

[0089] ＜補記＞

本発明の装置は、例えば単一のハードウェアエンティティとして、キーボードなどが接続可能な入力部、液晶ディスプレイなどが接続可能な出力部、ハードウェアエンティティの外部に通信可能な通信装置（例えば通信ケーブル

ル)が接続可能な通信部、CPU (Central Processing Unit、キャッシュメモリやレジスタなどを備えていてもよい)、メモリであるRAMやROM、ハードディスクである外部記憶装置並びにこれらの入力部、出力部、通信部、CPU、RAM、ROM、外部記憶装置の間のデータのやり取りが可能なように接続するバスを有している。また必要に応じて、ハードウェアエンティティに、CD-ROMなどの記録媒体を読み書きできる装置(ドライブ)などを設けることとしてもよい。このようなハードウェア資源を備えた物理的実体としては、汎用コンピュータなどがある。

[0090] ハードウェアエンティティの外部記憶装置には、上述の機能を実現するために必要となるプログラムおよびこのプログラムの処理において必要となるデータなどが記憶されている(外部記憶装置に限らず、例えばプログラムを読み出し専用記憶装置であるROMに記憶しておくこととしてもよい)。また、これらのプログラムの処理によって得られるデータなどは、RAMや外部記憶装置などに適宜に記憶される。

[0091] ハードウェアエンティティでは、外部記憶装置(あるいはROMなど)に記憶された各プログラムとこの各プログラムの処理に必要なデータが必要に応じてメモリに読み込まれて、適宜にCPUで解釈実行・処理される。その結果、CPUが所定の機能(上記、…部、…手段などと表した各構成要件)を実現する。

[0092] 本発明は上述の実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲で適宜変更が可能である。また、上記実施形態において説明した処理は、記載の順に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されるとしてもよい。

[0093] 既述のように、上記実施形態において説明したハードウェアエンティティ(本発明の装置)における処理機能をコンピュータによって実現する場合、ハードウェアエンティティが有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより

、上記ハードウェアエンティティにおける処理機能がコンピュータ上で実現される。

[0094] 上述の各種の処理は、図20に示すコンピュータ10000の記録部10020に、上記方法の各ステップを実行させるプログラムを読み込ませ、制御部10010、入力部10030、出力部10040などに動作させることで実施できる。図20の例に限らず、制御部10010をサーバ装置に配置し、入力部10030をクライアント装置に配置し、クライアント装置で入力された内容をサーバ装置が受け取り、判定して処理した結果をクライアント装置に返すという、サーバ・クライアント型で実施してもよい。

[0095] この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等のようなものでもよい。具体的には、例えば、磁気記録装置として、ハードディスク装置、フレキシブルディスク、磁気テープ等を、光ディスクとして、DVD (Digital Versatile Disc)、DVD-RAM (Random Access Memory)、CD-ROM (Compact Disc Read Only Memory)、CD-R (Recordable) / RW (ReWritable) 等を、光磁気記録媒体として、MO (Magneto-Optical disc) 等を、半導体メモリとしてEEPROM (Electrically Erasable and Programmable-Read Only Memory) 等を用いることができる。

[0096] また、このプログラムの流通は、例えば、そのプログラムを記録したDVD、CD-ROM等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。

[0097] このようなプログラムを実行するコンピュータは、例えば、まず、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送され

たプログラムを、一旦、自己の記憶装置に格納する。そして、処理の実行時、このコンピュータは、自己の記録媒体に格納されたプログラムを読み取り、読み取ったプログラムに従った処理を実行する。また、このプログラムの別の実行形態として、コンピュータが可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるASP（Application Service Provider）型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

[0098] また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、ハードウェアエンティティを構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

請求の範囲

[請求項1]

文書の一部を本文として抽出する本文抽出部と、
前記文書のうち前記本文に含まれない一文を不要文として選択する不要文選択部と、
前記本文に前記不要文を挿入して疑似問題を生成する不要文挿入部と、
前記不要文を選択肢とし、前記選択肢の前後に存在する文章のつながりが自然であることを意味する第1のラベルを前記疑似問題に付して、第1の学習用データを生成する第1学習用データ生成部と、
前記本文のうちの一文を選択肢とし、前記選択肢の前後に存在する文章のつながりが不自然であることを意味する第2のラベルを前記疑似問題に付して、第2の学習用データを生成する第2学習用データ生成部を含む
学習用データ生成装置。

[請求項2]

文書の一部を本文として抽出する本文抽出部と、
前記文書のうち前記本文に含まれない一文を不要文として選択する不要文選択部と、
前記本文に前記不要文を挿入して疑似問題を生成する不要文挿入部と、
所定の入力フォーマットに従った文章と、前記文章のつながりが自然であるか不自然であるかのラベルを含む学習用データを生成する学習用データ生成部を含み、
前記入力フォーマットは、
前記疑似問題のうちの一文を選択肢として、前記選択肢を除く前記疑似問題全体と前記選択肢とを連結した第1の入力フォーマット、または
前記疑似問題の各文を各選択肢とし、各選択肢から選びだした一つの選択肢と、選び出されなかったすべての選択肢からなる選択肢群と

を連結した第2の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの
前記選択肢よりも前に存在する文と、前記選択肢と、前記疑似問題の
うちの前記選択肢よりも後に存在する文とを連結した第3の入力フォ
ーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの
前記選択肢よりも前に存在する文と、前記疑似問題のうちの前記選
択肢よりも後に存在する文とを連結した第4の入力フォーマット、また
は

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの
前記選択肢よりも前に存在するN個（Nは1以上の整数）の文と、前
記選択肢および前記選択肢よりも後に存在するN個の文を含む文とを
連結した入力フォーマットAと、前記疑似問題のうちの前記選択肢よ
りも前に存在するN個（Nは1以上の整数）の文と、前記選択肢よ
りも後に存在するN+1個の文を含む文とを連結した入力フォーマット
Bの双方からなる第5の入力フォーマット、のいずれかである

学習用データ生成装置。

[請求項3]

請求項1または2に記載の学習用データ生成装置であって、
前記不要文選択部は、

前記文書のうちの前記本文に含まれない一文のうち、前記本文との
文間類似度が高い一文を前記不要文として選択する

学習用データ生成装置。

[請求項4]

請求項1から3の何れかに記載の学習用データ生成装置であって、
前記不要文挿入部は、

前記本文のうちの隣接する文のペアのうち、前記不要文との文間類
似度が高いペアの間に前記不要文を挿入する

学習用データ生成装置。

[請求項5]

所定の入力フォーマットに従った文章と、前記文章のつながりが自

然であるか不自然であるかのラベルを含む学習用データを用いて学習された判定器と、

判定対象の文章から前記所定の入力フォーマットに従った判定用データを作成して前記判定器に入力する入力フォーマット作成部と、

前記判定器から前記判定対象の文章のつながりが自然であるか否かの尤度を取得する尤度取得部と、

前記尤度に基づいて前記判定対象の文章のつながりが自然であるか否かを判定する判定部を含み、

前記入力フォーマットは、

文書の一部を本文とし、前記文書のうち前記本文に含まれない一文である不要文を前記本文に挿入して生成した疑似問題のうちの一文を選択肢として、前記選択肢を除く前記疑似問題全体と前記選択肢とを連結した第1の入力フォーマット、または

前記疑似問題の各文を各選択肢とし、各選択肢から選びだした一つの選択肢と、選び出されなかったすべての選択肢からなる選択肢群とを連結した第2の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの前記選択肢よりも前に存在する文と、前記選択肢と、前記疑似問題のうちの前記選択肢よりも後に存在する文とを連結した第3の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの前記選択肢よりも前に存在する文と、前記疑似問題のうちの前記選択肢よりも後に存在する文とを連結した第4の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの前記選択肢よりも前に存在するN個（Nは1以上の整数）の文と、前記選択肢および前記選択肢よりも後に存在するN個の文を含む文とを連結した入力フォーマットAと、前記疑似問題のうちの前記選択肢よ

りも前に存在する N 個（ N は1以上の整数）の文と、前記選択肢よりも後に存在する $N+1$ 個の文を含む文とを連結した入力フォーマット B の双方からなる第5の入力フォーマット、のいずれかである判定装置。

[請求項6]

文書の一部を本文として抽出するステップと、
前記文書のうち前記本文に含まれない一文を不要文として選択する不要文選択部と、
前記本文に前記不要文を挿入して疑似問題を生成するステップと、
前記不要文を選択肢とし、前記選択肢の前後に存在する文章のつながりが自然であることを意味する第1のラベルを前記疑似問題に付して、第1の学習用データを生成するステップと、
前記本文のうちの一文を選択肢とし、前記選択肢の前後に存在する文章のつながりが不自然であることを意味する第2のラベルを前記疑似問題に付して、第2の学習用データを生成するステップを含む学習用データ生成方法。

[請求項7]

所定の入力フォーマットに従った文章と、前記文章のつながりが自然であるか不自然であるかのラベルを含む学習用データを用いて学習された判定器を含む判定装置が実行する判定方法であって、
判定対象の文章から前記所定の入力フォーマットに従った判定用データを作成して前記判定器に入力するステップと、
前記判定器から前記判定対象の文章のつながりが自然であるか否かの尤度を取得するステップと、
前記尤度に基づいて前記判定対象の文章のつながりが自然であるか否かを判定するステップを含み、
前記入力フォーマットは、
文書の一部を本文とし、前記文書のうち前記本文に含まれない一文である不要文を前記本文に挿入して生成した疑似問題のうちの一文を選択肢として、前記選択肢を除く前記疑似問題全体と前記選択肢とを

連結した第1の入力フォーマット、または

前記疑似問題の各文を各選択肢とし、各選択肢から選びだした一つの選択肢と、選び出されなかったすべての選択肢からなる選択肢群とを連結した第2の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの前記選択肢よりも前に存在する文と、前記選択肢と、前記疑似問題のうちの前記選択肢よりも後に存在する文とを連結した第3の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの前記選択肢よりも前に存在する文と、前記疑似問題のうちの前記選択肢よりも後に存在する文とを連結した第4の入力フォーマット、または

前記疑似問題のうちの一文を選択肢として、前記疑似問題のうちの前記選択肢よりも前に存在するN個（Nは1以上の整数）の文と、前記選択肢および前記選択肢よりも後に存在するN個の文を含む文とを連結した入力フォーマットAと、前記疑似問題のうちの前記選択肢よりも前に存在するN個（Nは1以上の整数）の文と、前記選択肢よりも後に存在するN+1個の文を含む文とを連結した入力フォーマットBの双方からなる第5の入力フォーマット、のいずれかである

判定方法。

[請求項8] コンピュータを請求項1から5の何れかに記載の装置として機能させるプログラム。

[図1]

図1

本文＋不要文（二疑似問題）

おじいさんがその竹を切ってみると、なんと中には
 小さな女の子がいたのです。中でも特に熱心な若者
 が、五人いました。子どももないおじいさんとお
 ばあさんは、とても喜びました。そしてその子を『
 かぐやひめ』と名付けて、大切に育てたのです。か
 ぐやひめは大きくなるにしたがって、とても美しく
 なりました。



文書

おじいさんがその竹を切ってみると、なんと中には
 小さな女の子がいたのです。子どももないおじい
 さんとおばあさんは、とても喜びました。そしてそ
 の子を『かぐやひめ』と名付けて、大切に育てたの
 です。かぐやひめは大きくなるにしたがって、とて
 も美しくなりました。そして年頃になると、「どう
 か、かぐやひめをお嫁さんにください」と、若者が
 たくさんやってきました。中でも特に熱心な若者が
 、五人いました。みんな…（後略）

[図2]

図2

文書

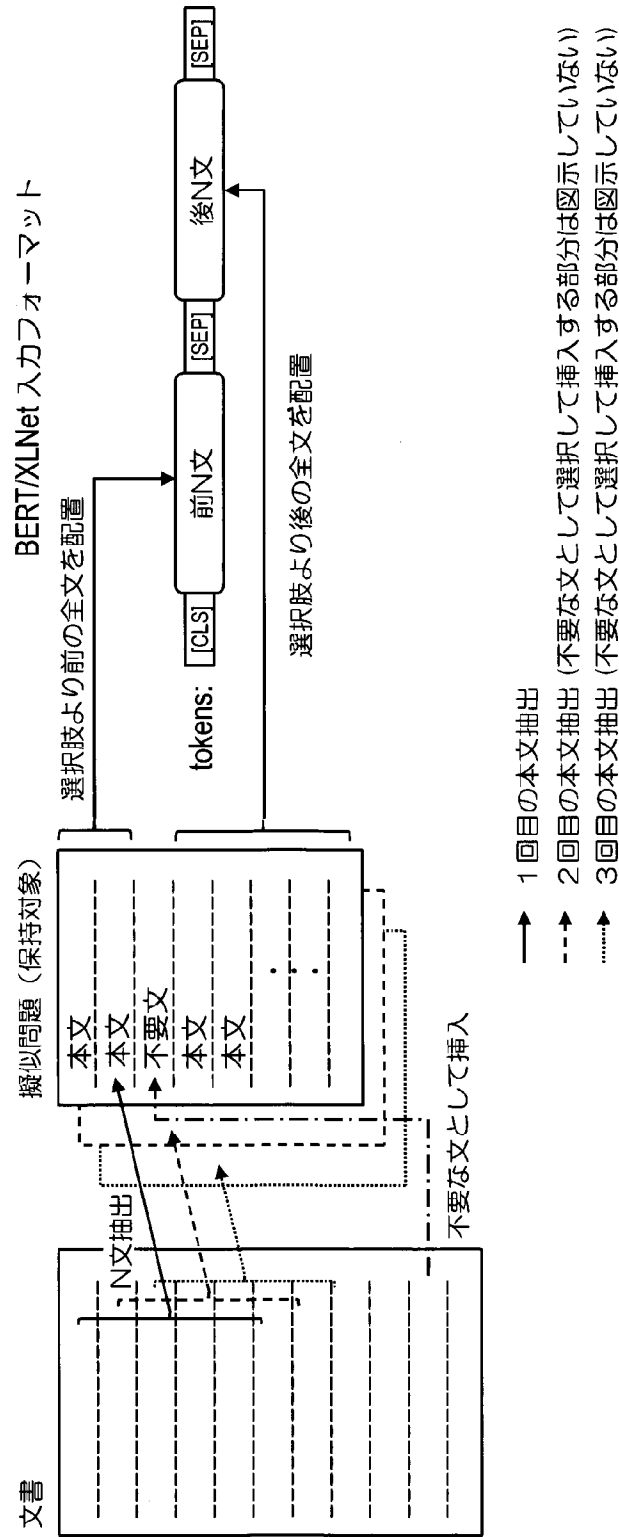
おじいさんがその竹を切ってみると、なんと中には
 小さな女の子がいたのです。子どものおじい
 さんとおばあさんは、とても喜びました。そしてそ
 の子を『かぐやひめ』と名付けて、大切に育てたの
 です。かぐやひめは大きくなるに連れて、とて
 も美しくなりました。そして年頃になると、「どう
 か、かぐやひめをお嫁さんにください」と、若者が
 たくさんやってきました。中でも特に熱心な若者が
 、五人いました。みんな…（中略）…
 なんとか五人の若者を追い返したかぐやひめですが
 、かぐやひめのうわさはとうとうみかどの耳にも入
 りました。
 …（後略）



本文＋不要文（＝疑似問題）

子どものおじいさんとおばあさんは、とても
 喜びました。そしてその子を『かぐやひめ』と名付
 けて、大切に育てたのです。かぐやひめは大きくな
 るに連れて、とても美しくなりました。なんと
 か五人の若者を追い返したかぐやひめですが、かぐ
 やひめのうわさはとうとうみかどの耳にも入りまし
 た。そして年頃になると、「どうか、かぐやひめを
 お嫁さんにください」と、若者がたくさんやってきました。

[図3]



[図3]

[図4]

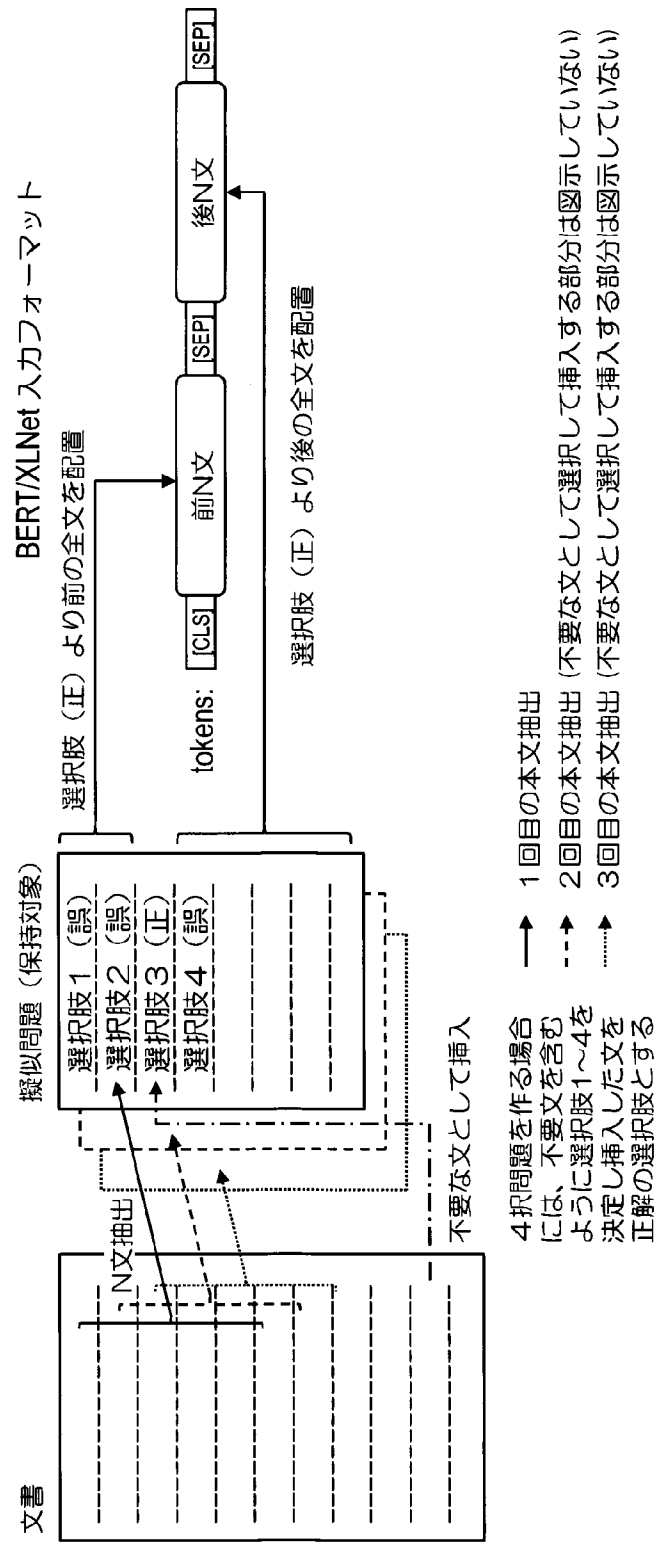
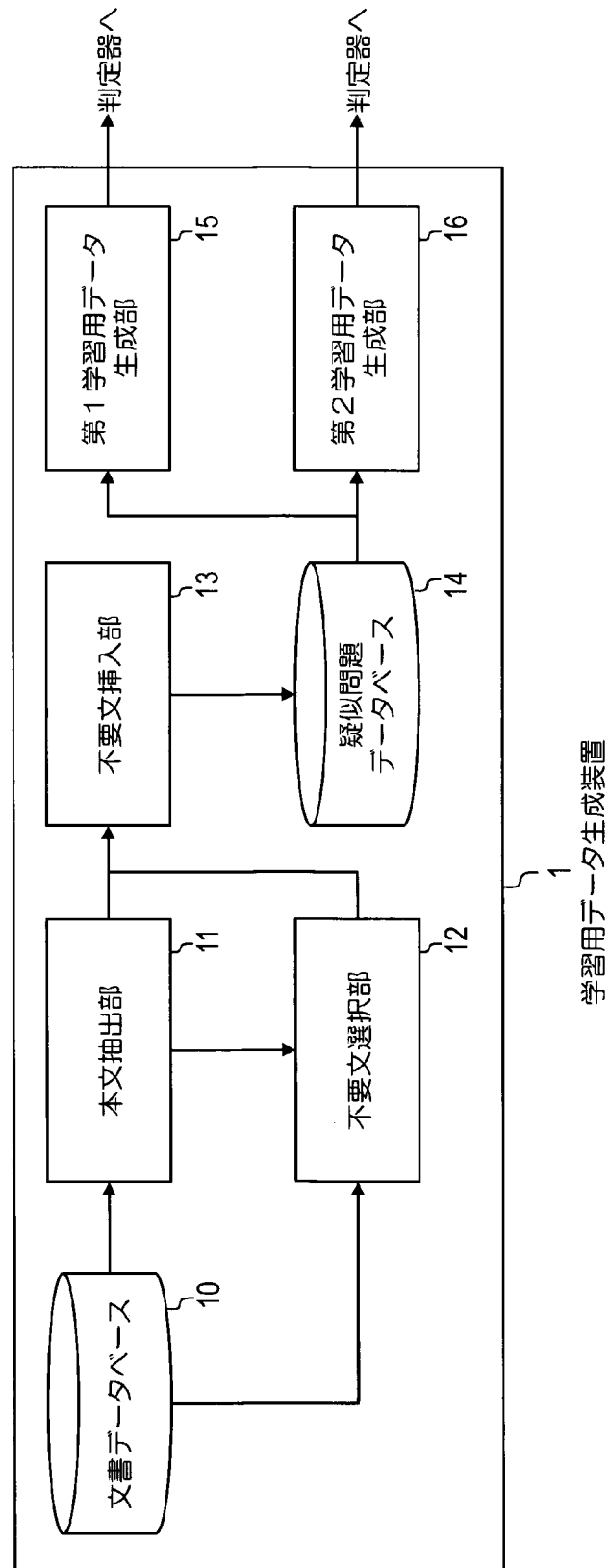
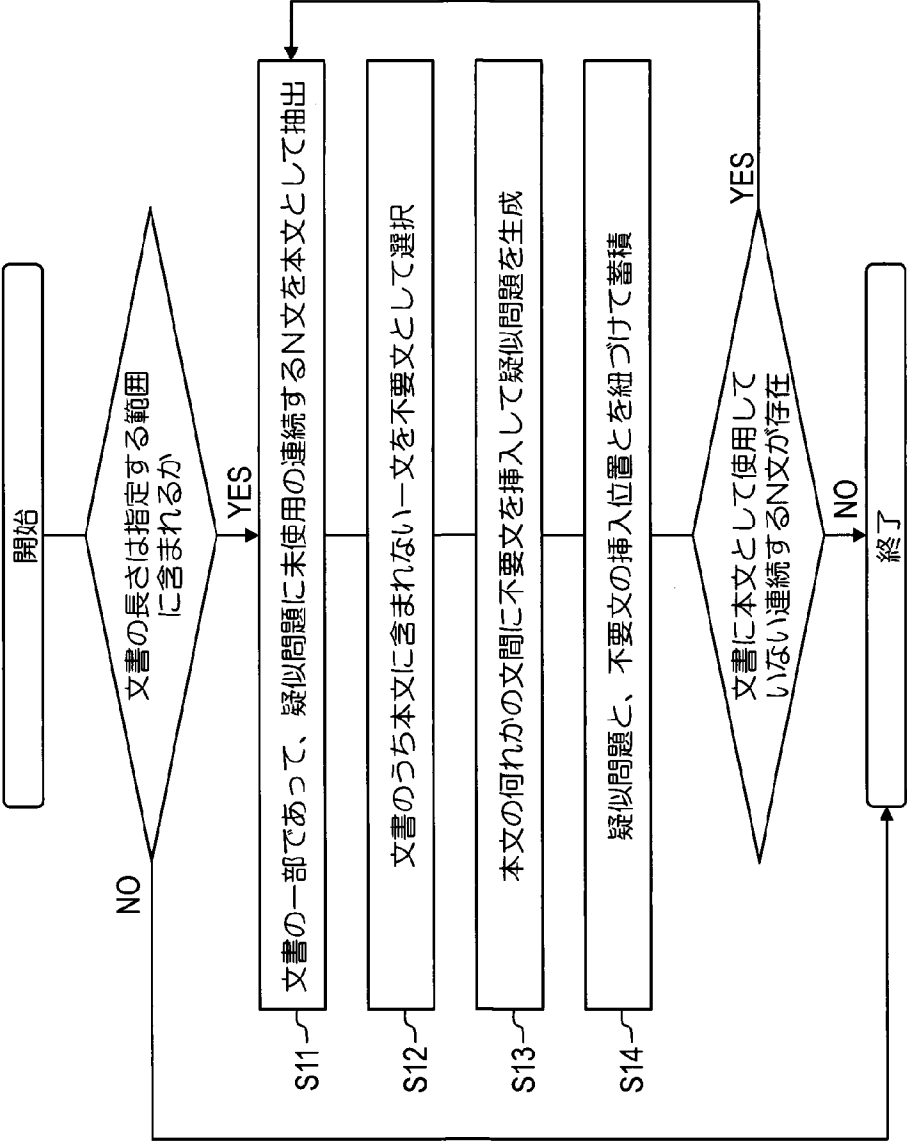


図4

[図5]



[図6]



※処理が終了したら次の文書で開始

図6

[図7]

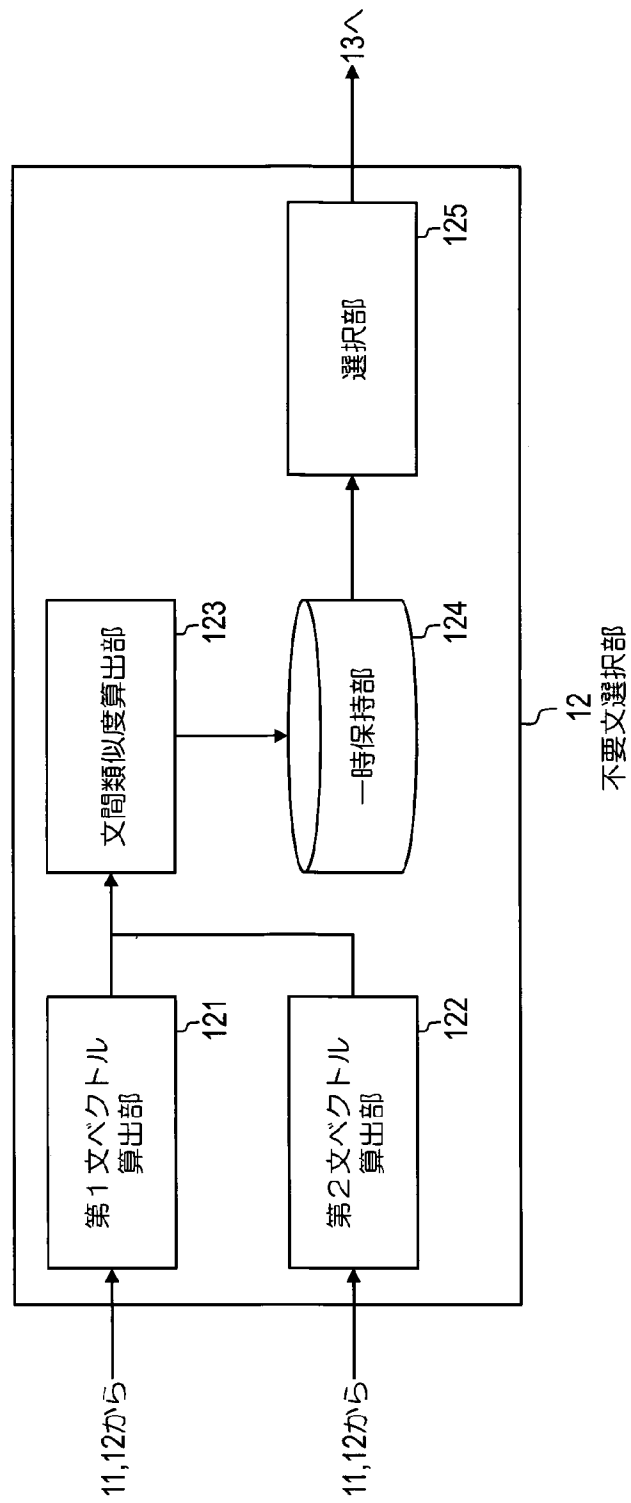
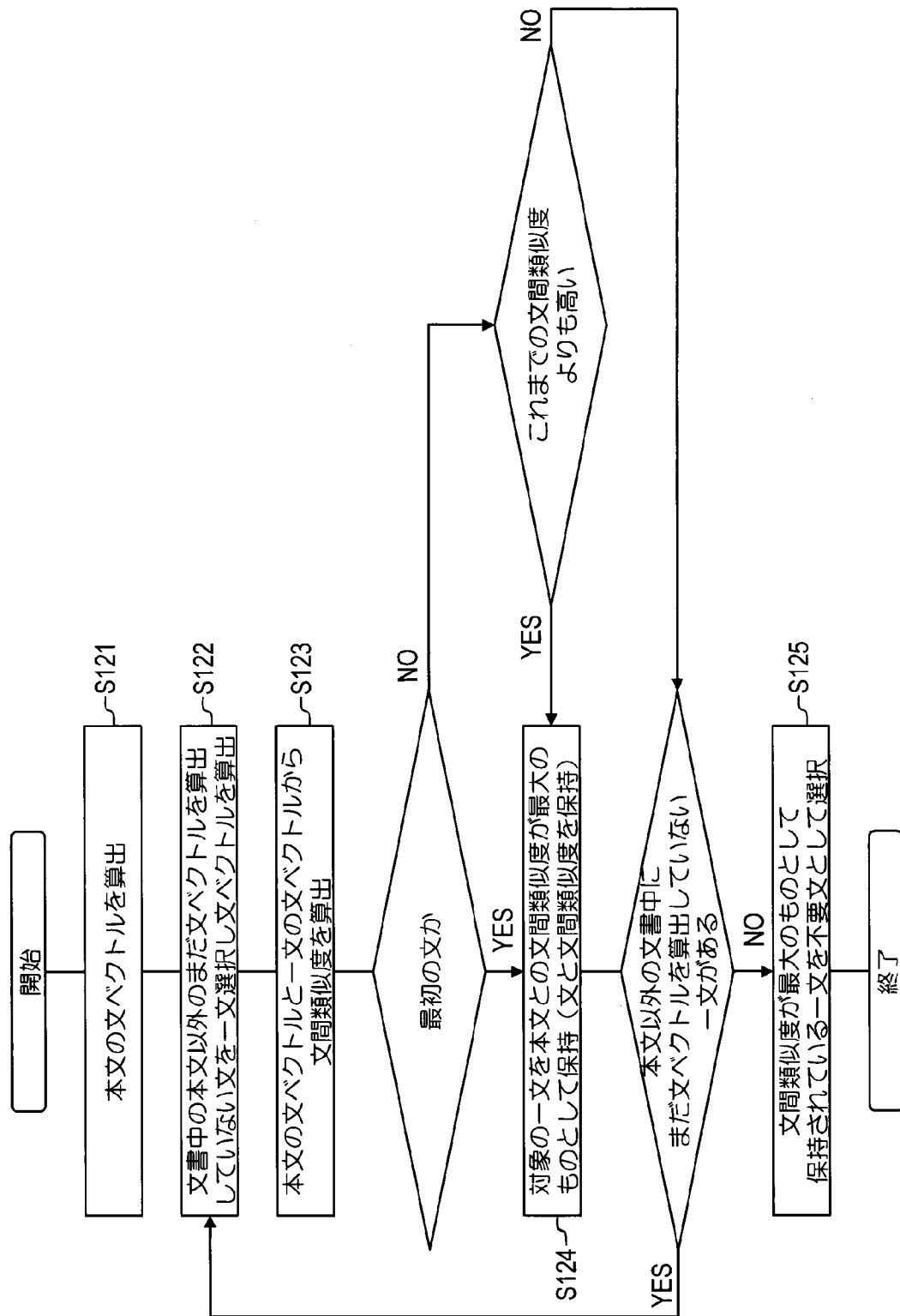
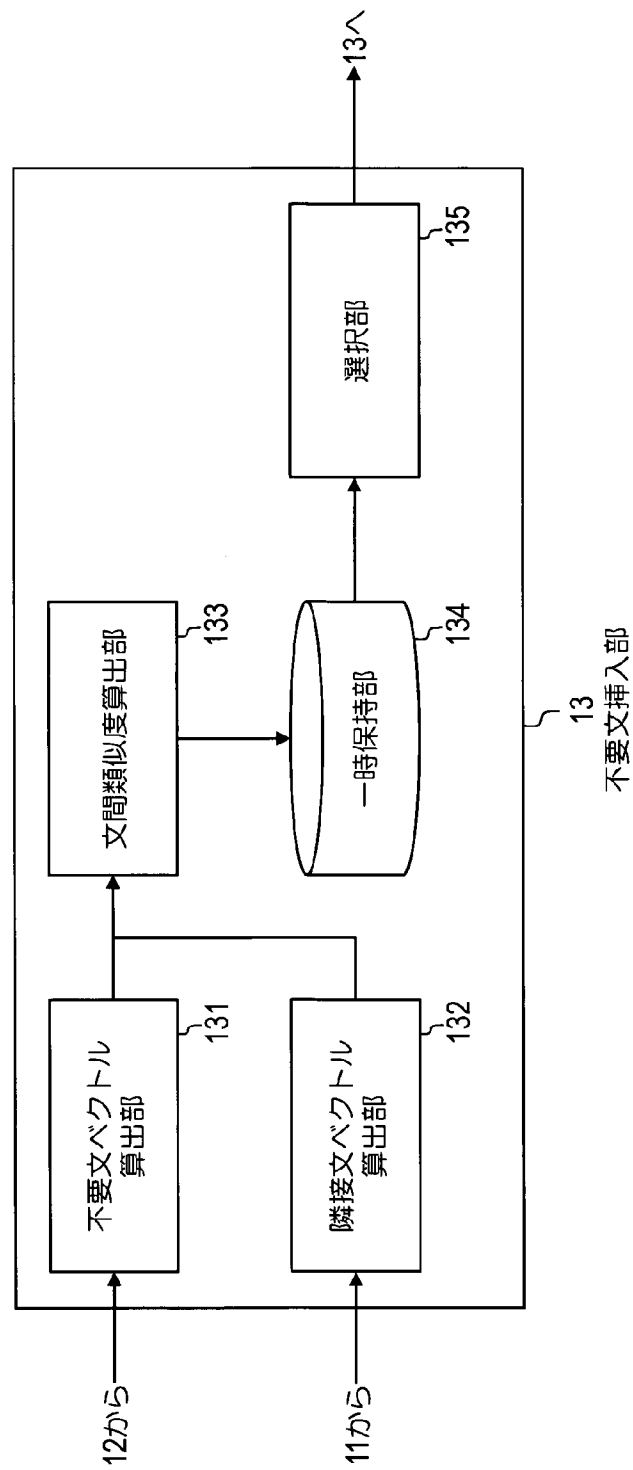


図7

[図8]



[図9]



[図10]

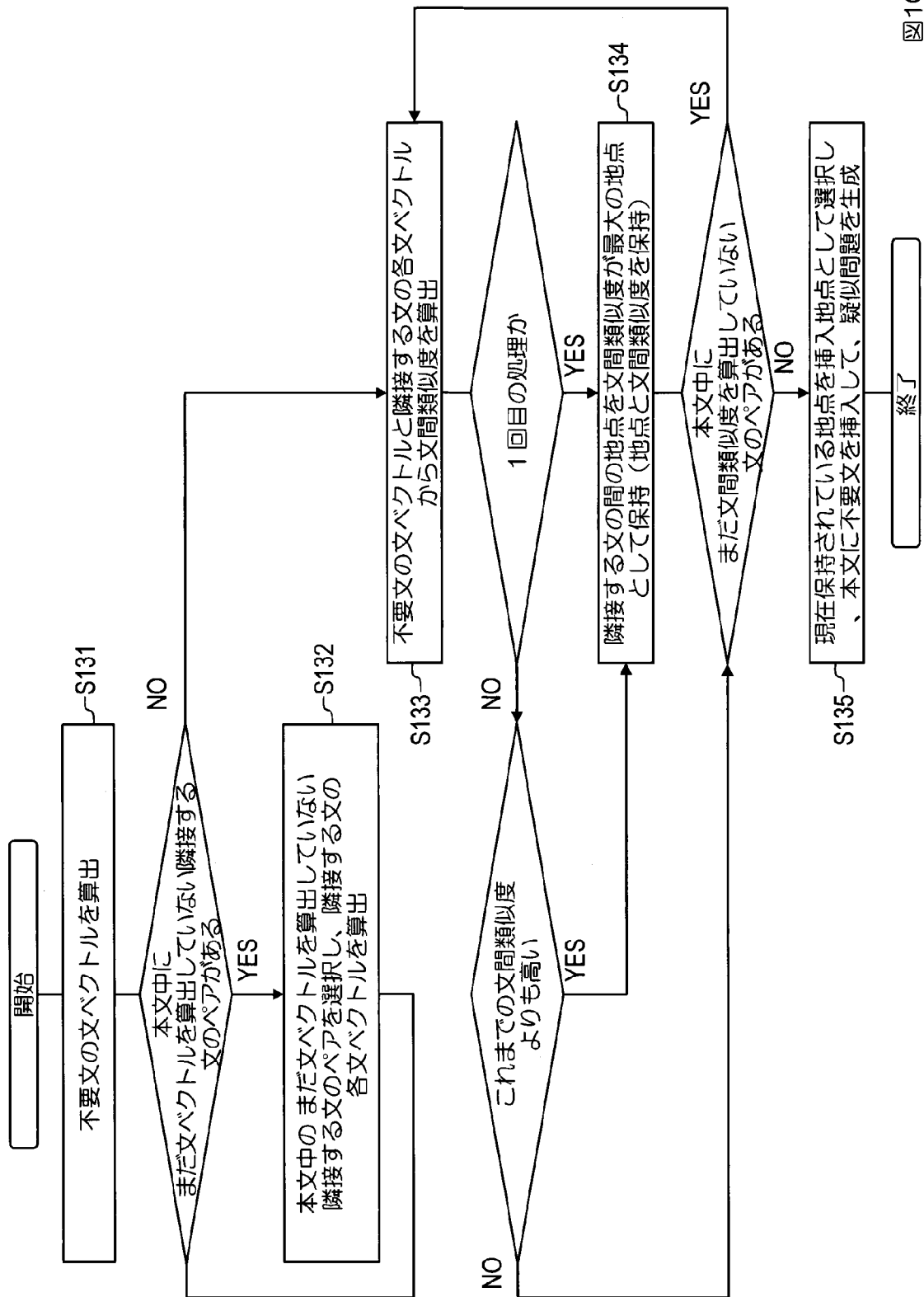
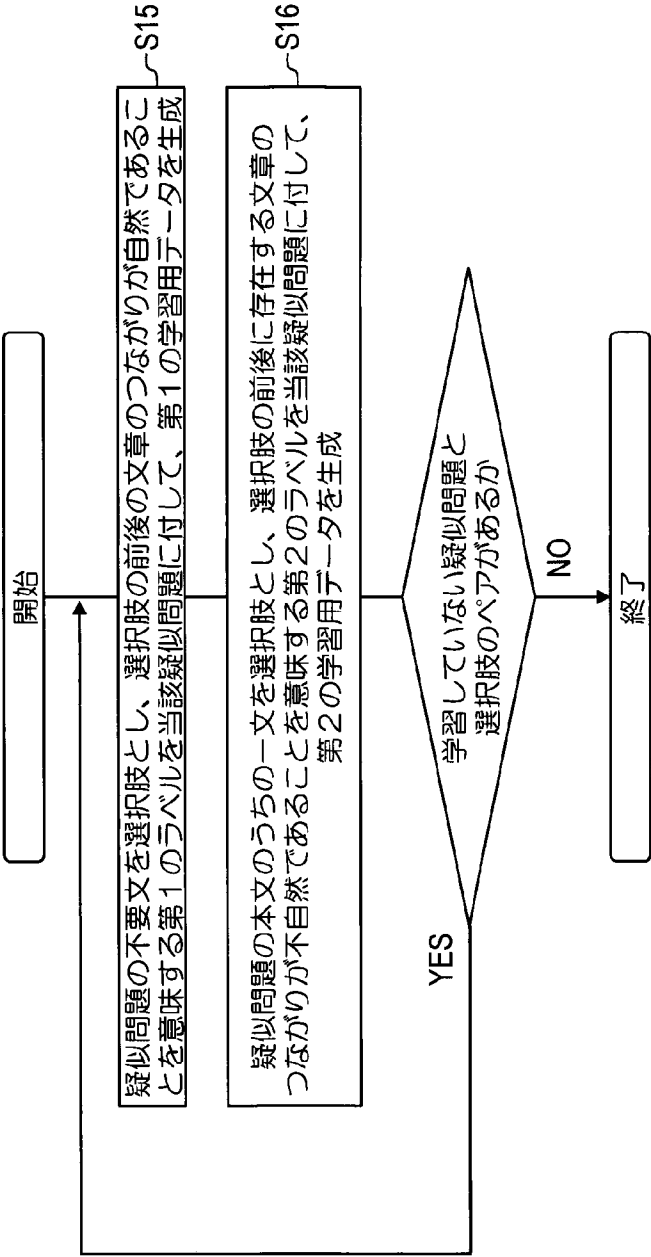


図10

[図11]



[図12]

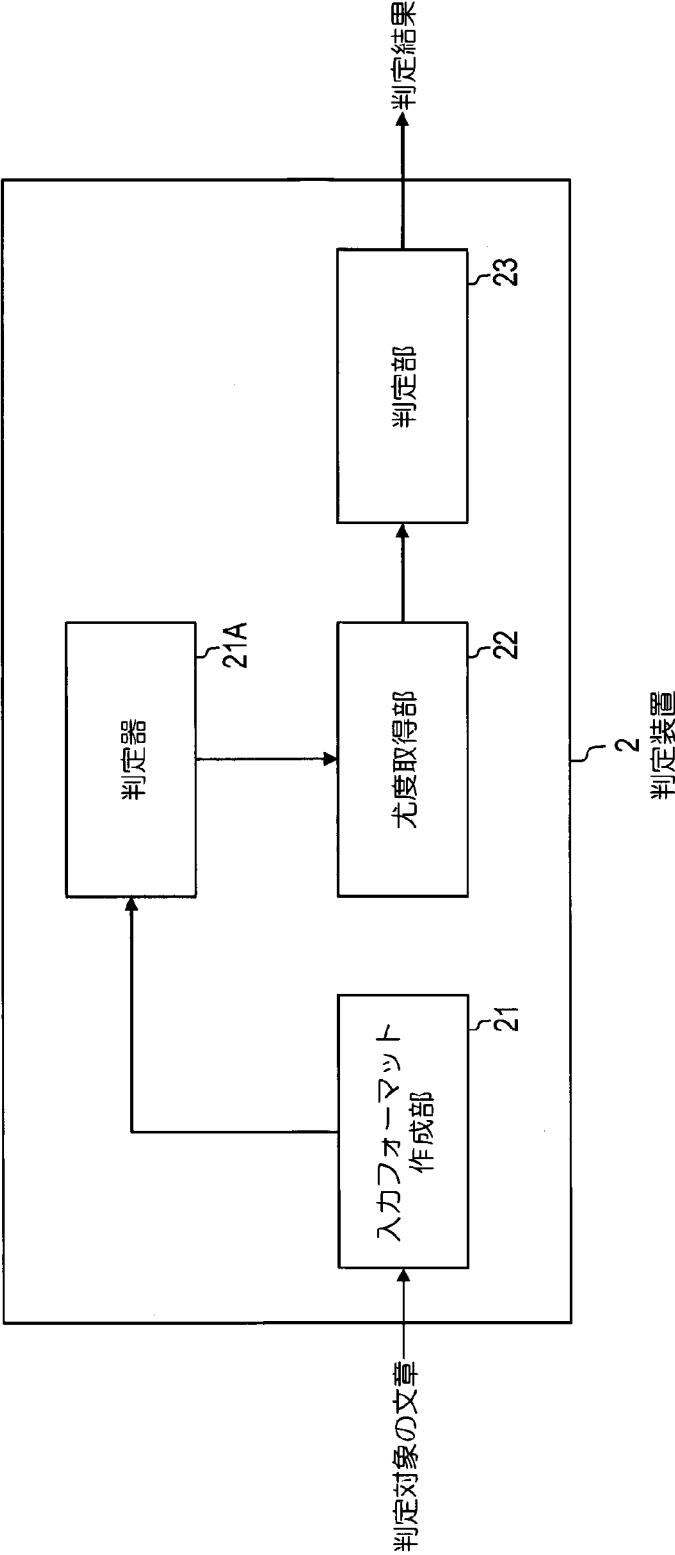


図12

[図13]

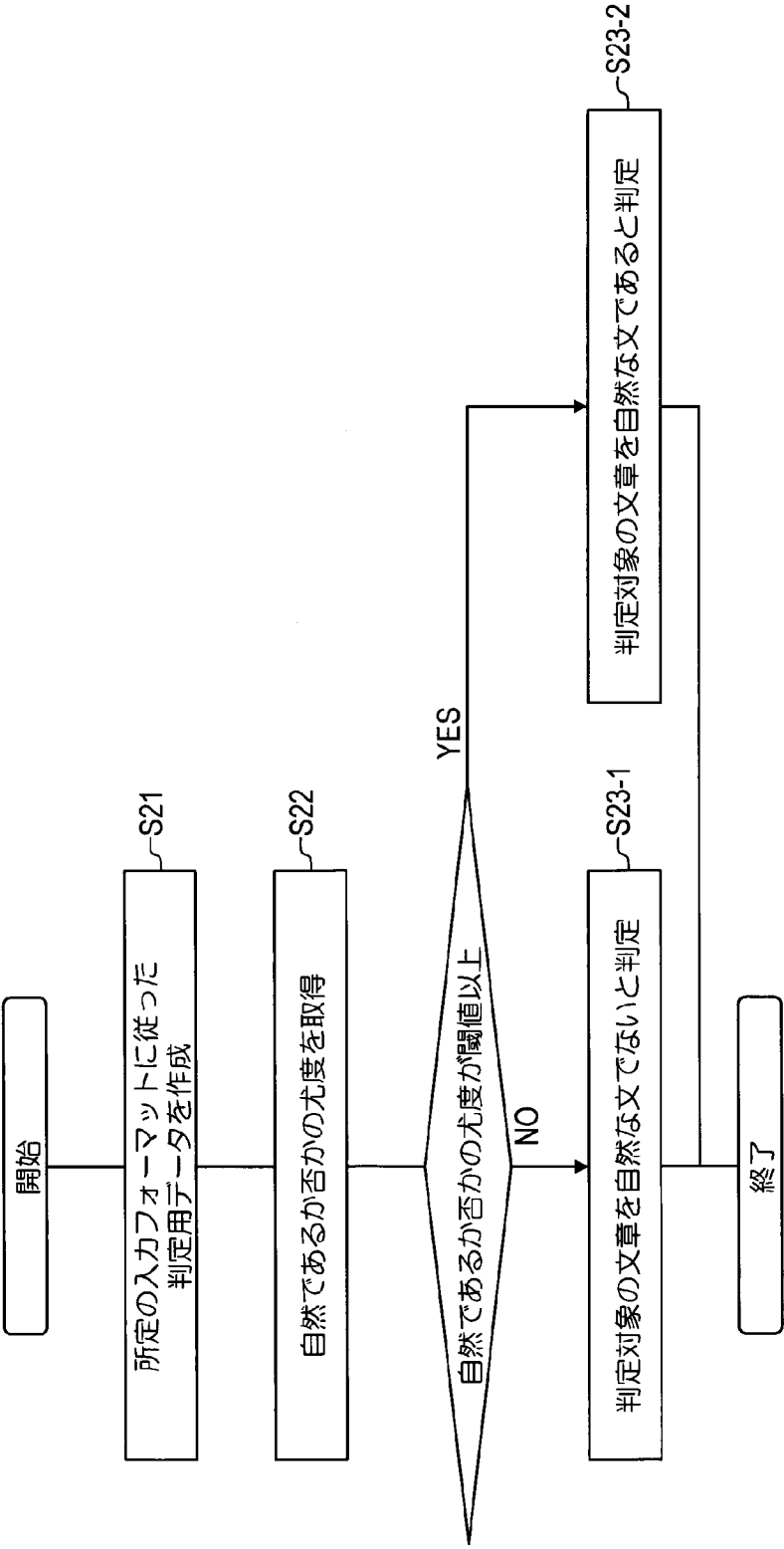


図13

[図14]

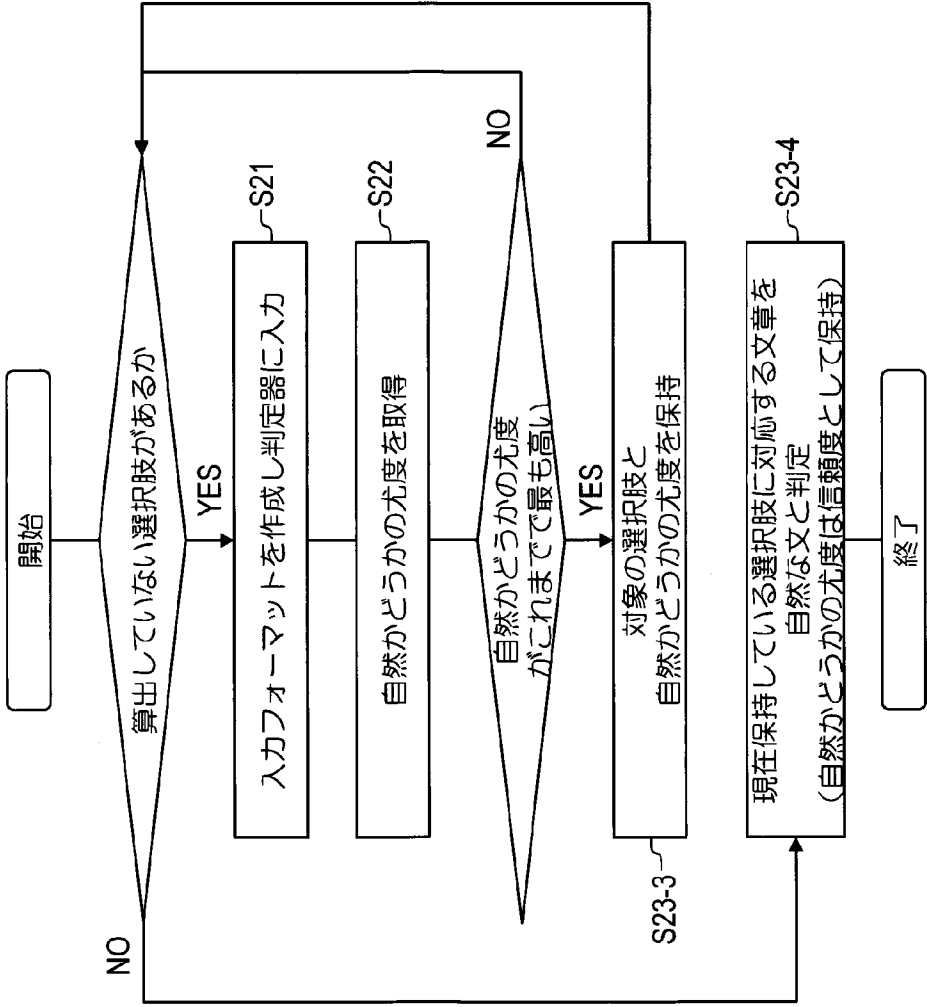


図14

[図15]

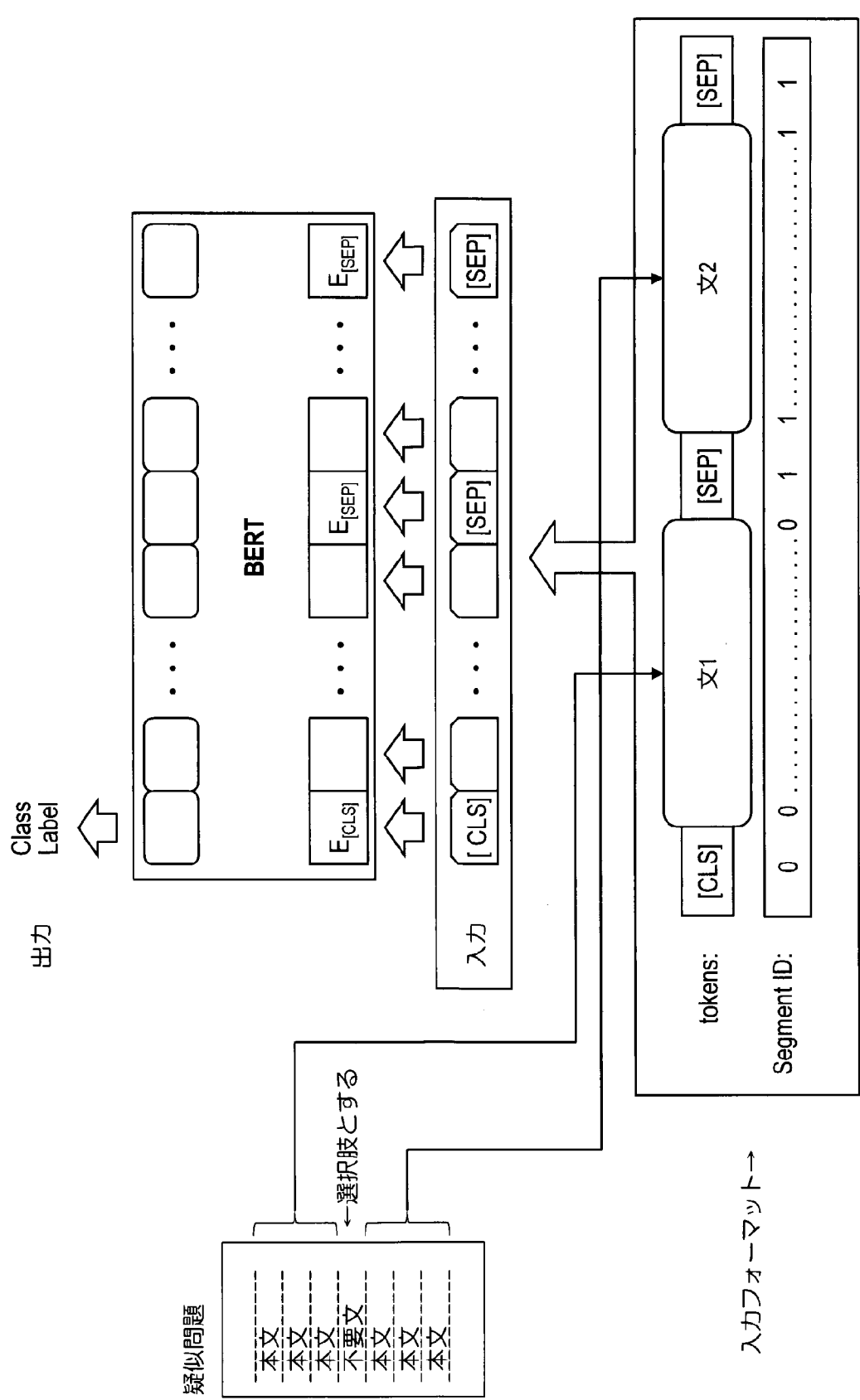
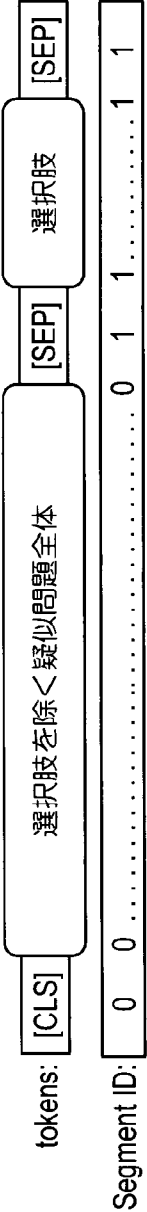


図15

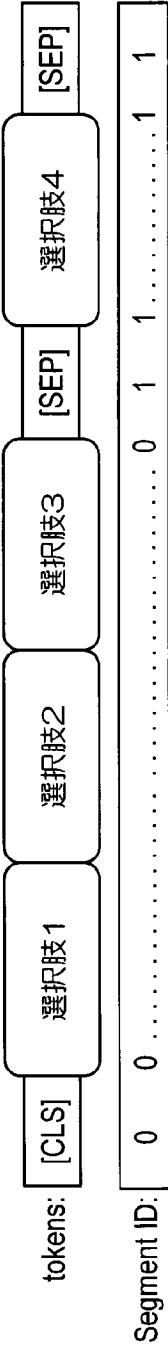
[図16]

入力フォーマット(1)
【疑似問題全体と選択肢の関連性 (疑似問題全体-選択肢)】



→ 文書全体と選択肢の類似性や含意関係が学習され、無関係な文に対して、不要を判断できる

入力フォーマット(2)
【選択肢間の関連性 (選択肢-選択肢)】

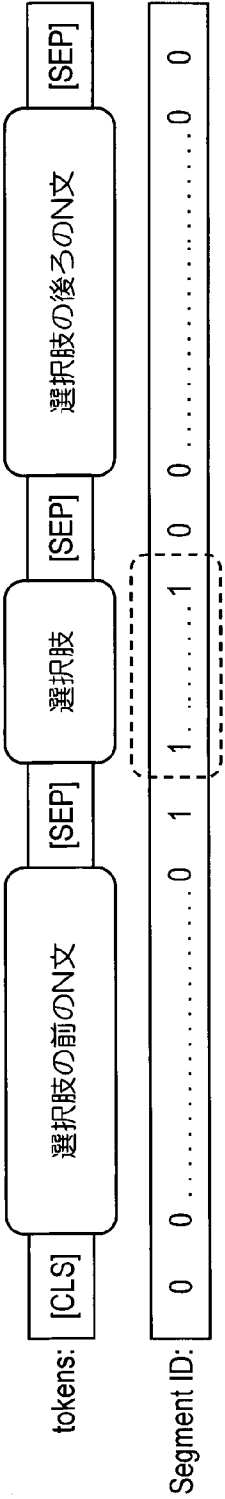


→ 文書全体ではなく、選択肢のみに限定し比較することで、選択肢間の違いを分類する学習が効果的に行える

図16

[図17]

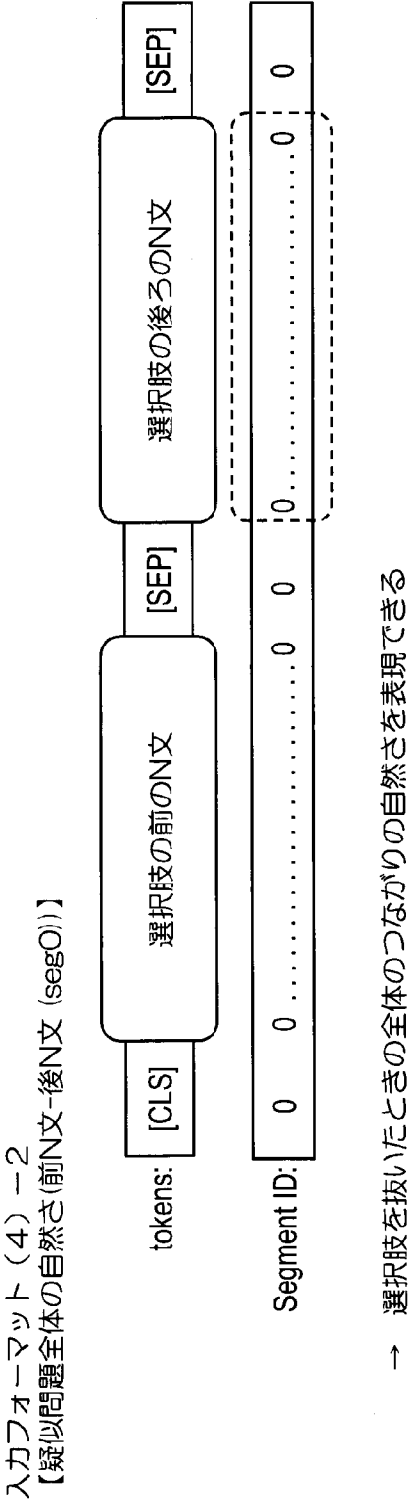
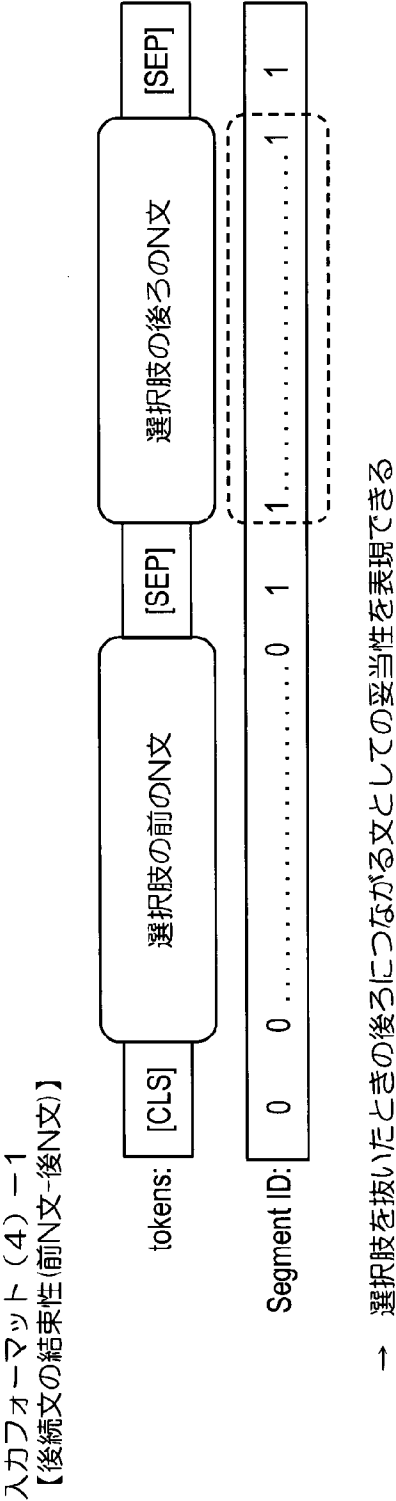
入力フォーマット (3)
【選択枝の不要らしさ (前N文-選択枝-後N文)】



→ 選択枝前後のつながりの自然さを評価できる

図17

[図18]

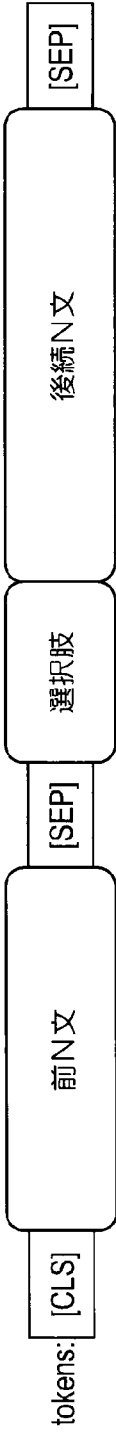


[図19]

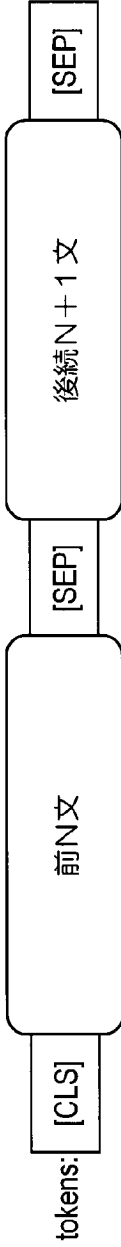
入力フォーマット (5)
【言語モデルの尤度差 (転移学習なし)】

以下2つの尤度の差が大きい選択肢を答えとして選択

A: 選択肢を含めたとき(以下を埋め込んだ場合)の尤度 S_a



B: 選択肢を抜いたとき(以下を埋め込んだ場合)の尤度 S_b



尤度差 $|S_a - S_b|$ が大きいときその選択肢を不要とする

図19

[図20]

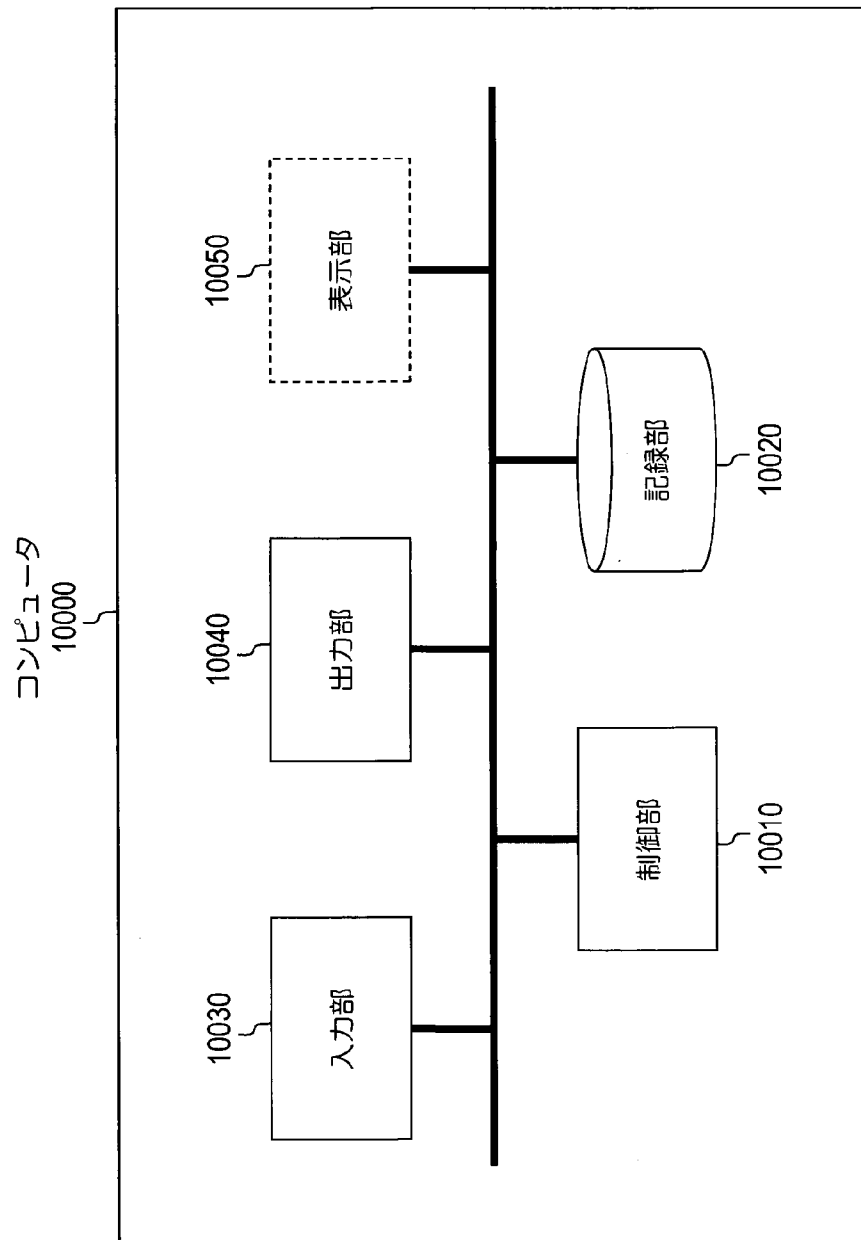


図20

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/020334

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G06F40/253 (2020.01) i

FI: G06F40/253

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. 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 1971-2020

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 2009-140411 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 25 June 2009, entire text, all drawings	1-8
A	笹野遼平他 1 名. 文脈解析 - 述語項構造・照応・談話構造の解析-. First Edition, Second Printing, Corona Publishing Co., Ltd., 05 February 2018, section 4.6, non-official translation (SASANO, Ryohei, IIDA, Ryu. Context Analysis: Analysis of Predicate-Argument Structure, Anaphora, Discourse Structure.)	1-8
A	CN 111078892 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 28 April 2020, entire text, all drawings	1-8



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

01.10.2020

Date of mailing of the international search report

13.10.2020

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/JP2020/020334

Patent Documents referred to in the Report	Publication Date	Patent Family	Publication Date
JP 2009-140411 A	25.06.2009	(Family: none)	
CN 111078892 A	28.04.2020	(Family: none)	

A. 発明の属する分野の分類（国際特許分類（IPC）） G06F 40/253(2020.01)i FI: G06F40/253										
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 <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 - 2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2020年</td> </tr> </table> 国際調査で使用了電子データベース（データベースの名称、調査に使用した用語）			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
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									
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 2009-140411 A（日本電信電話株式会社）25.06.2009（2009 - 06 - 25） 全文・全図	1-8								
A	笹野遼平 他1名，文脈解析－述語項構造・照応・談話構造の解析－，初版第2刷，株式会社コロナ社，2018.02.05 4,6節	1-8								
A	CN 111078892 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 28.04.2020（2020 - 04 - 28） 全文・全図	1-8								
☐ 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 01.10.2020		国際調査報告の発送日 13.10.2020								
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） 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/JP2020/020334

引用文献			公表日	パテントファミリー文献	公表日
JP	2009-140411	A	25.06.2009	(ファミリーなし)	
CN	111078892	A	28.04.2020	(ファミリーなし)	