(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)公開番号
特開2022-51113
(P2022-51113A)
(43)公開日　令和4年3月31日(2022.3.31)

(51)国際特許分類		FI		テーマコード（参考）
Ｇ０６Ｆ　40/216 (2020.01)		Ｇ０６Ｆ	40/216	５Ｂ０９１
Ｇ０６Ｆ　40/56　(2020.01)		Ｇ０６Ｆ	40/56	

審査請求　未請求　請求項の数　10　　ＯＬ　（全14頁）

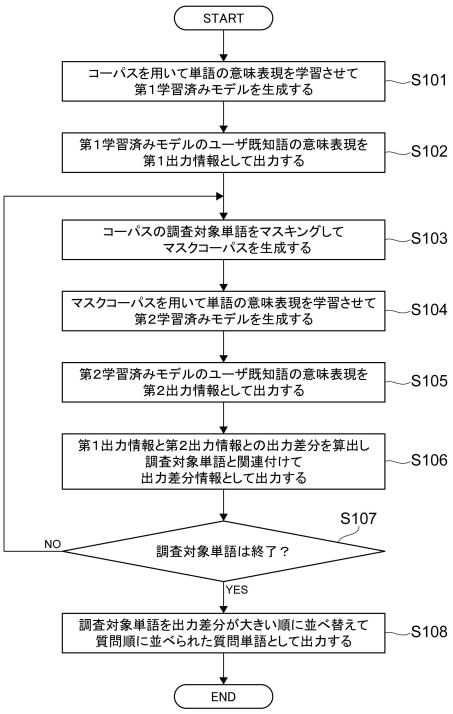
(21)出願番号　特願2020-157394(P2020-157394)
(22)出願日　令和2年9月18日(2020.9.18)

(71)出願人　000005496
富士フイルムビジネスイノベーション株式会社
東京都港区赤坂九丁目７番３号
(74)代理人　100104880
弁理士　古部 次郎
(74)代理人　100118108
弁理士　久保 洋之
(72)発明者　稲木 誓哉
神奈川県横浜市西区みなとみらい六丁目１番 富士ゼロックス株式会社内

Ｆターム（参考）　5B091　EA01

(54)【発明の名称】　情報出力装置、質問生成装置、及びプログラム

(57)【要約】　（修正有）【課題】アンケート等でユーザ既知語を調べる場合に比較して、ユーザ既知語を効率的に調べることを可能とする情報出力装置、質問生成装置及びプログラムを提供する。
【解決手段】質問生成装置は、プロセッサを備える。プロセッサは、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた特定のユーザ既知語の意味表現との差分を算出し、差分に基づいて、対象単語がユーザ既知語である可能性に関する情報を出力する。
【選択図】図７



【特許請求の範囲】
【請求項１】
プロセッサを備え、
前記プロセッサは、
特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、当該特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた当該特定のユーザ既知語の意味表現との差分を算出し、
前記差分に基づいて、前記対象単語がユーザ既知語である可能性に関する情報を出力することを特徴とする情報出力装置。
【請求項２】
前記プロセッサは、
複数の対象単語の各対象単語について前記差分を算出することにより、複数の差分を算出し、
前記対象単語がユーザ既知語である可能性に関する情報として、前記複数の対象単語の各対象単語についての前記差分に基づく順序で並べられた当該複数の対象単語を出力することを特徴とする請求項１に記載の情報出力装置。
【請求項３】
前記差分に基づく順序は、当該差分が大きい順序であることを特徴とする請求項２に記載の情報出力装置。
【請求項４】
前記第２のモデルは、前記特定の例文集合の前記対象単語を除く部分を用いて単語の意味表現を未学習モデルに新たに学習させたモデルであることを特徴とする請求項１に記載の情報出力装置。
【請求項５】
前記特定の例文集合の前記対象単語を除く部分は、当該特定の例文集合の前記特定のユーザ既知語及び当該対象単語の少なくとも何れか一方を含む構成要素の当該対象単語を除く部分であることを特徴とする請求項４に記載の情報出力装置。
【請求項６】
前記第２のモデルは、前記特定の例文集合の前記対象単語を除く部分を用いて単語の意味表現を学習済みモデルに更に学習させたモデルであることを特徴とする請求項１に記載の情報出力装置。
【請求項７】
前記特定の例文集合の前記対象単語を除く部分は、当該特定の例文集合の当該対象単語を含む構成要素の当該対象単語を除く部分であることを特徴とする請求項６に記載の情報出力装置。
【請求項８】
プロセッサを備え、
前記プロセッサは、
複数の対象単語の各対象単語について、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、当該特定の例文集合の当該各対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた当該特定のユーザ既知語の意味表現との差分を算出することにより、複数の差分を算出し、
前記複数の差分に基づいて、前記複数の対象単語を用いた質問を生成する
ことを特徴とする質問生成装置。
【請求項９】
前記プロセッサは、
前記特定のユーザ既知語に代えて、前記質問に対するユーザの回答から把握される他のユーザ既知語を用いて、前記複数の差分を算出し、

前記複数の差分に基づいて、前記複数の対象単語を用いた質問を再生成することを特徴とする請求項８に記載の質問生成装置。
【請求項１０】
コンピュータに、
特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、当該特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた当該特定のユーザ既知語の意味表現との差分を算出する機能と、
前記差分に基づいて、前記対象単語がユーザ既知語である可能性に関する情報を出力する機能と
を実現させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、情報出力装置、質問生成装置、及びプログラムに関する。
【背景技術】
【０００２】
あるパターンが他のパターンを含意しているような２つの言語パターンのペアを生成する技術は、知られている（例えば、特許文献１参照）。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特許第６５５１９６８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
例えば、ユーザに単語を用いた質問を行い、その質問に対するユーザの回答を利用して何らかの処理を行うことがある。その際、ユーザがその単語の意味を知らないと、回答の質又は量が低下するので、その単語はユーザが意味を知っている単語（以下、「ユーザ既知語」という）であることが望ましい。ここで、ユーザ既知語を調べるには、アンケート等を行うことも考えられるが、時間やコスト等の観点から効率的な方法ではない。
【０００５】
本発明の目的は、アンケート等でユーザ既知語を調べる場合に比較して、ユーザ既知語を効率的に調べることを可能とすることにある。
【課題を解決するための手段】
【０００６】
請求項１に記載の発明は、プロセッサを備え、前記プロセッサは、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、当該特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた当該特定のユーザ既知語の意味表現との差分を算出し、前記差分に基づいて、前記対象単語がユーザ既知語である可能性に関する情報を出力することを特徴とする情報出力装置である。
請求項２に記載の発明は、前記プロセッサは、複数の対象単語の各対象単語について前記差分を算出することにより、複数の差分を算出し、前記対象単語がユーザ既知語である可能性に関する情報として、前記複数の対象単語の各対象単語についての前記差分に基づく順序で並べられた当該複数の対象単語を出力することを特徴とする請求項１に記載の情報出力装置である。
請求項３に記載の発明は、前記差分に基づく順序は、当該差分が大きい順序であることを特徴とする請求項２に記載の情報出力装置である。
請求項４に記載の発明は、前記第２のモデルは、前記特定の例文集合の前記対象単語を除

く部分を用いて単語の意味表現を未学習モデルに新たに学習させたモデルであることを特徴とする請求項１に記載の情報出力装置である。
請求項５に記載の発明は、前記特定の例文集合の前記対象単語を除く部分は、当該特定の例文集合の前記特定のユーザ既知語及び当該対象単語の少なくとも何れか一方を含む構成要素の当該対象単語を除く部分であることを特徴とする請求項４に記載の情報出力装置である。
請求項６に記載の発明は、前記第２のモデルは、前記特定の例文集合の前記対象単語を除く部分を用いて単語の意味表現を学習済みモデルに更に学習させたモデルであることを特徴とする請求項１に記載の情報出力装置である。
請求項７に記載の発明は、前記特定の例文集合の前記対象単語を除く部分は、当該特定の例文集合の当該対象単語を含む構成要素の当該対象単語を除く部分であることを特徴とする請求項６に記載の情報出力装置である。
請求項８に記載の発明は、プロセッサを備え、前記プロセッサは、複数の対象単語の各対象単語について、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、当該特定の例文集合の当該各対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた当該特定のユーザ既知語の意味表現との差分を算出することにより、複数の差分を算出し、前記複数の差分に基づいて、前記複数の対象単語を用いた質問を生成することを特徴とする質問生成装置である。
請求項９に記載の発明は、前記プロセッサは、前記特定のユーザ既知語に代えて、前記質問に対するユーザの回答から把握される他のユーザ既知語を用いて、前記複数の差分を算出し、前記複数の差分に基づいて、前記複数の対象単語を用いた質問を再生成することを特徴とする請求項８に記載の質問生成装置である。
請求項１０に記載の発明は、コンピュータに、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、当該特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた当該特定のユーザ既知語の意味表現との差分を算出する機能と、前記差分に基づいて、前記対象単語がユーザ既知語である可能性に関する情報を出力する機能とを実現させるためのプログラムである。
【発明の効果】
【０００７】
請求項１の発明によれば、アンケート等でユーザ既知語を調べる場合に比較して、ユーザ既知語を効率的に調べることが可能となる。
請求項２の発明によれば、複数の対象単語についてのユーザ既知語である可能性に基づく順序を知ることができる。
請求項３の発明によれば、複数の対象単語についてのユーザ既知語である可能性が高い順序を知ることができる。
請求項４の発明によれば、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させるモデルとして学習済みモデルを用意しなくても、ユーザ既知語を調べることが可能となる。
請求項５の発明によれば、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させるモデルとして未学習モデルを用意した場合において、特定の例文集合を削減することができる。
請求項６の発明によれば、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させるモデルとして未学習モデルを用意する場合に比較して、単語の意味表現を学習させたモデルの精度を向上することができる。
請求項７の発明によれば、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させるモデルとして学習済みモデルを用意した場合において、特定の例文集合を削減することができる。
請求項８の発明によれば、アンケート等でユーザ既知語を調べてユーザ既知語を用いた質

問を生成する場合に比較して、ユーザ既知語を用いた質問を効率的に生成することが可能となる。
請求項９の発明によれば、特定のユーザ既知語のみを用いて質問を生成する場合に比較して、ユーザ既知語を用いた質問が生成される可能性が高まる。
請求項１０の発明によれば、アンケート等でユーザ既知語を調べる場合に比較して、ユーザ既知語を効率的に調べることが可能となる。
【図面の簡単な説明】
【０００８】
【図１】本発明の実施の形態における質問生成装置のハードウェア構成例を示した図である。
【図２】本発明の実施の形態における質問生成装置の機能構成例を示したブロック図である。
【図３】（ａ），（ｂ）は本発明の実施の形態における質問生成装置で記憶されるコーパスの具体例を示した図である。
【図４】（ａ），（ｂ）は本発明の実施の形態における質問生成装置で記憶される学習済みモデルの具体例を示した図である。
【図５】（ａ），（ｂ）は本発明の実施の形態における質問生成装置で記憶される出力情報の具体例を示した図である。
【図６】本発明の実施の形態における質問生成装置で記憶される出力差分情報の具体例を示した図である。
【図７】本発明の実施の形態における質問生成装置の動作例を示したフローチャートである。
【発明を実施するための形態】
【０００９】
以下、添付図面を参照して、本発明の実施の形態について詳細に説明する。
【００１０】
［本実施の形態の概要］
本実施の形態は、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた特定のユーザ既知語の意味表現との差分を算出し、その差分に基づいて、対象単語がユーザ既知語である可能性に関する情報を出力する情報出力装置である。
【００１１】
ここで、情報出力装置は、１つの対象単語について差分を算出し、この差分が閾値以上であれば、対象単語がユーザ既知語である可能性に関する情報として、対象単語がユーザ既知語と判断される旨の情報を出力するものでもよい。
【００１２】
或いは、情報出力装置は、複数の対象単語の各対象単語について差分を算出することにより、複数の差分を算出し、対象単語がユーザ既知語である可能性に関する情報として、複数の対象単語の各対象単語についての差分に基づく順序で並べられた複数の対象単語を出力するものでもよい。
【００１３】
情報出力装置は、これらの何れであってもよいが、以下では、後者であるものとして説明する。そして、単に複数の対象単語を出力するのではなく、複数の対象単語を用いた質問を生成するものとする。
【００１４】
その場合、本実施の形態は、複数の対象単語の各対象単語について、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、特定の例文集合の各対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた特定のユーザ既知語の意味表現との差分を算出することにより、複

数の差分を算出し、複数の差分に基づいて、複数の対象単語を用いた質問を生成する質問生成装置となる。

【００１５】

従って、以下では、本実施の形態が質問生成装置である場合を例にとって説明する。

【００１６】

ここで、質問生成装置とは、ユーザに与える質問を生成する装置である。この装置は、例えば、質問に対するユーザの回答を利用して、目的のタスクを解くシステムにおいて、質問を生成する装置であってよい。タスクとしては、例えば、単語分類、単語間関連度予測がある。

【００１７】

システムが質問を与える方法としては、次のような方法が考えられる。

【００１８】

目的のタスクが単語分類タスクである場合は、システムが、単語と分類項目とを提示して、その単語に最も関連しそうな分類項目をユーザに質問する、という方法である。

【００１９】

目的のタスクが単語間関連度予測タスクである場合は、システムが、２つの単語を提示して、それらがどのくらい関連しているかをユーザに質問する、という方法である。

【００２０】

また、例文集合とは、何らかの例文を集めたものをいう。例文は、記事や書物等の一般に「文書」と呼ばれ得る比較的長い文であってもよいし、会話の文等の比較的短い文であってもよい。また、例文は、テキストデータとして記録された文だけでなく、例えば、音声データとして記録された文を含んでもよい。更に、例文は、自然言語処理の研究目的に限らず、如何なる目的で集められたものでもよい。以下では、例文集合としてコーパスを例にとって説明する。

【００２１】

更に、特定の例文集合の対象単語を除く部分とは、特定の例文集合に何らかの処理を行って対象単語が含まないようにされた部分のことをいう。この処理は、例えば、対象単語をマスクする処理でもよいし、対象単語を一時的に削除する処理でもよいが、以下では、前者の処理を例にとって説明する。

【００２２】

更に、単語の意味表現とは、単語の意味をベクトル化して表現したものをいう。但し、本実施の形態では、単語の意味表現により単語の意味の近さを計算できればよいので、単語の意味表現は、少なくとも単語の意味の近さを計算できる他の方法で表現したものであってもよい。

【００２３】

更にまた、差分に基づく順序とは、差分を用いて定められる順序をいう。差分に基づく順序は、例えば、差分が大きい順序でもよいし、差分が大きい順序を基本としつつ他の要素を加味した順序でもよい。ここで、他の要素は、他の複数のユーザ既知語を用いた場合の差分であってよい。例えば、特定のユーザ既知語のみを用いた場合の差分は小さいが、他の複数のユーザ既知語を用いた場合の差分の平均が大きい場合や分散が小さい場合に、順序を上げることが考えられる。或いは、他の要素は、対象単語の文法上の属性等であってもよい。以下では、差分に基づく順序として差分が大きい順序を用いた場合を例にとって説明する。

【００２４】

［質問生成装置のハードウェア構成］

図１は、本実施の形態における質問生成装置１０のハードウェア構成例を示した図である。図示するように、質問生成装置１０は、演算手段であるプロセッサ１１と、記憶手段であるメインメモリ１２及びＨＤＤ（Hard Disk Drive）１３とを備える。ここで、プロセッサ１１は、ＯＳ（Operating System）やアプリケーション等の各種ソフトウェアを実行し、後述する各機能を実現する。また、メインメモリ１２は、各種ソフトウェアや

その実行に用いるデータ等を記憶する記憶領域であり、ＨＤＤ１３は、各種ソフトウェアに対する入力データや各種ソフトウェアからの出力データ等を記憶する記憶領域である。更に、質問生成装置１０は、外部との通信を行うための通信Ｉ／Ｆ（以下、「Ｉ／Ｆ」と表記する）１４と、ディスプレイ等の表示デバイス１５と、キーボードやマウス等の入力デバイス１６とを備える。

【００２５】

［質問生成装置の機能構成］

図２は、本実施の形態における質問生成装置１０の機能構成例を示したブロック図である。図示するように、質問生成装置１０は、コーパス記憶部２１と、第１学習部２２と、第１学習済みモデル記憶部２３と、第１出力部２４と、第１出力情報記憶部２５とを備えている。また、質問生成装置１０は、マスキング処理部３１を備えている。更に、質問生成装置１０は、マスクコーパス記憶部４１と、第２学習部４２と、第２学習済みモデル記憶部４３と、第２出力部４４と、第２出力情報記憶部４５とを備えている。更にまた、質問生成装置１０は、出力差分算出部５１と、出力差分情報記憶部５２と、ランキング処理部５３と、質問単語記憶部５４とを備えている。

【００２６】

コーパス記憶部２１は、コーパスを記憶する。コーパスは、例えば、質問を行う分野における特定のコーパスである。コーパス記憶部２１に記憶されたコーパスの具体例については後述する。

【００２７】

第１学習部２２は、コーパス記憶部２１に記憶されたコーパスを用いて単語の意味表現をモデルに学習させることにより第１学習済みモデルを生成する。本実施の形態では、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルの一例として、第１学習済みモデルを用いている。ここで、第１学習部２２は、コーパス記憶部２１に記憶されたコーパスを用いて、全く学習していないモデルを学習させることにより、第１学習済みモデルを生成してもよい。或いは、第１学習部２２は、コーパス記憶部２１に記憶されたコーパスを用いて、既に学習したモデルを更新することにより、第１学習済みモデルを生成してもよい。

【００２８】

第１学習済みモデル記憶部２３は、第１学習部２２が生成した第１学習済みモデルを記憶する。第１学習済みモデル記憶部２３に記憶された第１学習済みモデルの具体例については後述する。

【００２９】

第１出力部２４は、第１学習済みモデル記憶部２３に記憶された第１学習済みモデルから得られた特定のユーザ既知語の意味表現を第１出力情報として出力する。本実施の形態では、第１のモデルから得られた特定のユーザ既知語の意味表現の一例として、第１出力情報を用いている。

【００３０】

第１出力情報記憶部２５は、第１出力部２４が出力した第１出力情報を記憶する。第１出力情報記憶部２５に記憶された第１出力情報の具体例については後述する。

【００３１】

マスキング処理部３１は、コーパス記憶部２１に記憶されたコーパスに対し、特定のユーザ既知語に対する寄与を調べたい対象の単語（以下、「調査対象単語」という）をマスクするマスキング処理を行うことにより、マスクコーパスを作成する。本実施の形態では、対象単語の一例として、調査対象単語を用いており、特定の例文集合の対象単語を除く部分の一例として、マスクコーパスを用いている。

【００３２】

マスクコーパス記憶部４１は、マスキング処理部３１が作成したマスクコーパスを記憶する。マスクコーパス記憶部４１に記憶されたマスクコーパスの具体例については後述する。

【００３３】
第２学習部４２は、マスクコーパス記憶部４１に記憶されたマスクコーパスを用いて単語の意味表現をモデルに学習させることにより第２学習済みモデルを生成する。本実施の形態では、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルの一例として、第２学習済みモデルを用いている。ここで、第２学習部４２は、マスクコーパス記憶部４１に記憶されたマスクコーパスを用いて、全く学習していないモデルを学習させることにより、第２学習済みモデルを生成してもよい。この場合、第２学習済みモデルは、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を未学習モデルに新たに学習させたモデルの一例である。或いは、第２学習部４２は、マスクコーパス記憶部４１に記憶されたマスクコーパスを用いて、既に学習したモデルを更新することにより、第２学習済みモデルを生成してもよい。この場合、第２学習済みモデルは、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習済みモデルに更に学習させたモデルの一例である。
【００３４】
第２学習済みモデル記憶部４３は、第２学習部４２が取得した第２学習済みモデルを記憶する。第２学習済みモデル記憶部４３に記憶された第２学習済みモデルの具体例については後述する。
【００３５】
第２出力部４４は、第２学習済みモデル記憶部４３に記憶された第２学習済みモデルから得られた特定のユーザ既知語の意味表現を第２出力情報として出力する。本実施の形態では、第２のモデルから得られた特定のユーザ既知語の意味表現の一例として、第２出力情報を用いている。
【００３６】
第２出力情報記憶部４５は、第２出力部４４が出力した第２出力情報を記憶する。第２出力情報記憶部４５に記憶された第２出力情報の具体例については後述する。
【００３７】
出力差分算出部５１は、複数の調査対象単語のそれぞれについて、第１出力情報記憶部２５に記憶された第１出力情報と、その調査対象単語を選択した場合に第２出力情報記憶部４５に記憶された第２出力情報との差分である出力差分を算出する。本実施の形態では、第１のモデルから得られた特定のユーザ既知語の意味表現と、第２のモデルから得られた特定のユーザ既知語の意味表現との差分を算出する手段の一例として、出力差分算出部５１を設けている。また、本実施の形態では、複数の対象単語の各対象単語について、第１のモデルから得られた特定のユーザ既知語の意味表現と、第２のモデルから得られた特定のユーザ既知語の意味表現との差分を算出することにより、複数の差分を算出する手段の一例としても、出力差分算出部５１を設けている。
【００３８】
出力差分情報記憶部５２は、複数の調査対象単語のそれぞれについて、その調査対象単語と、その調査対象単語を選択した場合に出力差分算出部５１が算出した出力差分とを関連付けた出力差分情報を記憶する。
【００３９】
ランキング処理部５３は、複数の調査対象単語を、ユーザに与える質問で用いる単語（以下、「質問単語」という）として、出力差分情報記憶部５２に記憶された出力差分が大きい順、つまり、ユーザ既知語である可能性が高い順に並べて出力する。これは、コーパス内に調査対象単語がある場合とない場合とで特定のユーザ既知語の意味表現が大きくずれるのであれば、調査対象単語がないとその特定のユーザ既知語の意味表現が得られないと考えられるので、調査対象単語はユーザ既知語と判断できる、という考え方に基づくものである。本実施の形態では、差分に基づいて、対象単語がユーザ既知語である可能性に関する情報を出力する手段の一例として、ランキング処理部５３を設けている。また、本実施の形態では、複数の差分に基づいて、複数の対象単語を用いた質問を生成する手段の一例としても、ランキング処理部５３を設けている。

【００４０】
質問単語記憶部５４は、ランキング処理部５３が出力した質問単語を、ランキング処理部５３が並べた順序で記憶する。そして、タスクを実行するシステムが、質問単語記憶部５４に記憶された質問単語を、質問単語記憶部５４に記憶された順序で取り出して、ユーザに与える質問で用いることになる。
【００４１】
尚、これらの機能部は、ソフトウェアとハードウェア資源とが協働することにより実現される。具体的には、これらの機能部は、プロセッサ１１が、これらを実現するプログラムを例えばＨＤＤ１３からメインメモリ１２に読み込んで実行することにより実現される。
【００４２】
次に、本実施の形態における質問生成装置１０で記憶されるコーパスの具体例について説明する。
【００４３】
図３（ａ）は、コーパス記憶部２１に記憶されるコーパスの具体例を示した図である。図示するように、コーパス記憶部２１に記憶されるコーパスは、文書２１１，２１２，２１３，…を含んでいる。そして、文書２１１は、文２１１１，２１１２，２１１３，…を含み、文書２１２は、文２１２１，２１２２，２１２３，…を含み、文書２１３は、文２１３１，２１３２，２１３３，…を含んでいる。ここで、ユーザ既知語ｎ１，ｎ２，ｎ３は、それぞれ、文２１１１，２１１３，２１３２に存在するものとする。
【００４４】
図３（ｂ）は、マスクコーパス記憶部４１に記憶されるマスクコーパスの具体例を示した図である。図示するように、マスクコーパス記憶部４１に記憶されるマスクコーパスは、コーパス記憶部２１に記憶されるコーパスにおいて調査対象単語がマスクされたものになっている。ここでは、調査対象単語ｍ１，ｍ２，ｍ３が、それぞれ、文４１１１，４１２３，４１３２に存在し、これらがマスクされているものとする。
【００４５】
ところで、図３（ａ），（ｂ）では、マスクコーパス記憶部４１に記憶されるデータの単位を文としたが、これには限らない。データの単位は、より一般化し、文書の構成要素としてよい。文書の構成要素には、文以外に、段落、章、節等が含まれる。
【００４６】
また、図３（ｂ）では、ユーザ既知語しか含まない文や、ユーザ既知語及び調査対象単語の何れも含まない文も、マスクコーパス記憶部４１に記憶したが、これには限らない。ユーザ既知語しか含まない文や、ユーザ既知語及び調査対象単語の何れも含まない文は、マスクコーパス記憶部４１に記憶しないようにしてもよい。
【００４７】
具体的には、第２学習部４２が、全く学習していないモデルを学習させる場合は、ユーザ既知語及び調査対象単語の何れかを含む文のみをフィルタリングして、マスクコーパス記憶部４１に記憶するとよい。つまり、図３（ｂ）の例で言えば、文４１１１，４１１３，４１２３，４１３２をマスクコーパス記憶部４１に記憶するとよい。これは、特定の例文集合の対象単語を除く部分が、特定の例文集合の特定のユーザ既知語及び対象単語の少なくとも何れか一方を含む構成要素の対象単語を除く部分である場合の一例である。
【００４８】
一方、第２学習部４２が、既に学習したモデルを更新する場合は、調査対象単語を含む文のみをフィルタリングして、マスクコーパス記憶部４１に記憶するとよい。つまり、図３（ｂ）の例で言えば、文４１１１，４１２３，４１３２をマスクコーパス記憶部４１に記憶するとよい。更新前の学習済みモデルにユーザ既知語が含まれていると仮定できるからである。これは、特定の例文集合の対象単語を除く部分が、特定の例文集合の対象単語を含む構成要素の対象単語を除く部分である場合の一例である。
【００４９】
次に、本実施の形態における質問生成装置１０で記憶される学習済みモデルの具体例につ

いて説明する。尚、以下では、Ｗｏｒｄ２Ｖｅｃを構成する２種類のモデルのうちＣＢＯＷ（Continuous Bag-Of-Words）モデルにより単語の意味表現を学習させる場合を例にとって説明する。

【００５０】

図４（ａ）は、第１学習済みモデル記憶部２３に記憶される第１学習済みモデルの具体例を示した図である。ここでは、コーパスＸを入力としたＣＢＯＷモデルの出力である第１学習済みモデルをＹと表記する。第１学習済みモデルＹは、単語の意味表現を各行に持つＶ×Ｗの行列である。Ｖは単語の数であり、Ｗは意味表現の次元数である。以下、第１学習済みモデルＹの単語ｖの行における次元ｗの意味表現をY_{v}（ｗ）と表すことにする。図において、第１学習済みモデルＹの１行目は、単語ｖ１の次元１，２，３，…の意味表現を表している。また、２行目は、単語ｖ２の次元１，２，３，…の意味表現を表し、３行目は、単語ｖ３の次元１，２，３，…における意味表現を表している。

【００５１】

図４（ｂ）は、第２学習済みモデル記憶部４３に記憶される第２学習済みモデルの具体例を示した図である。ここでは、マスキング処理部３１が調査対象単語ｍｊをマスキングしたコーパスＸをコーパスX^{mj}とし、このコーパスX^{mj}を入力としたＣＢＯＷモデルの出力である第２学習済みモデルをY^{mj}と表記する。第２学習済みモデルY^{mj}も、単語の意味表現を各行に持つＶ×Ｗの行列である。以下、第２学習済みモデルY^{mj}の単語ｖの行における次元ｗの意味表現をY^{mj}_{v}（ｗ）と表すことにする。図において、第２学習済みモデルY^{mj}の１行目は、単語ｖ１の次元１，２，３，…の意味表現を表している。また、２行目は、単語ｖ２の次元１，２，３，…の意味表現を表し、３行目は、単語ｖ３の次元１，２，３，…の意味表現を表している。

【００５２】

次に、本実施の形態における質問生成装置１０で記憶される出力情報の具体例について説明する。

【００５３】

図５（ａ）は、第１出力情報記憶部２５に記憶される第１出力情報の具体例を示した図である。図示するように、第１出力情報は、第１学習済みモデルＹからユーザ既知語ｎｉに対応する行を抜き出したものである。ここでは、この抜き出された行である第１出力情報をY_{ni}と表記する。第１出力情報Y_{ni}は、単語の意味表現を要素に持つＷ次元のベクトルである。

【００５４】

図５（ｂ）は、第２出力情報記憶部４５に記憶される第２出力情報の具体例を示した図である。図示するように、第２出力情報は、第２学習済みモデルY^{mj}からユーザ既知語ｎｉに対応する行を抜き出したものである。ここでは、この抜き出された行である第２出力情報をY^{mj}_{ni}と表記する。第２出力情報Y^{mj}_{ni}は、単語の意味表現を要素に持つＷ次元のベクトルである。

【００５５】

次に、本実施の形態における質問生成装置１０で記憶される出力差分情報の具体例について説明する。

【００５６】

図６は、出力差分情報記憶部５２に記憶される出力差分情報の具体例を示した図である。図示するように、出力差分情報は、調査対象単語と、出力差分とを対応付けたものである。調査対象単語はｍｊであり、出力差分はδ（ｎｉ，ｍｊ）である（ｊ＝１，２，３，…）。ここで、出力差分δ（ｎｉ，ｍｊ）は、第１出力情報Y_{ni}と、調査対象単語ｍｊをマスクした場合の第２出力情報Y^{mj}_{ni}との二乗距離として定義される。

【００５７】

尚、その後、ランキング処理部５３が、調査対象単語ｍｊを、出力差分δ（ｎｉ，ｍｊ）の大きい順に並べ替えて、質問単語記憶部５４に記憶することになる。

【００５８】

［質問生成装置の動作］
図７は、本実施の形態における質問生成装置１０の動作例を示したフローチャートである。
【００５９】
図示するように、質問生成装置１０では、まず、第１学習部２２が、コーパス記憶部２１に記憶されたコーパスを用いて単語の意味表現を学習させて第１学習済みモデルを生成する（ステップ１０１）。この第１学習済みモデルは、第１学習済みモデル記憶部２３に記憶される。
【００６０】
次に、第１出力部２４が、第１学習済みモデル記憶部２３に記憶された第１学習済みモデルからユーザ既知語の意味表現を抜き出して第１出力情報として出力する（ステップ１０２）。この第１出力情報は、第１出力情報記憶部２５に記憶される。
【００６１】
一方、質問生成装置１０では、マスキング処理部３１が、コーパス記憶部２１に記憶されたコーパスに対して調査対象単語をマスクするマスキング処理を行ってマスクコーパスを生成する（ステップ１０３）。このマスクコーパスは、マスクコーパス記憶部４１に記憶される。
【００６２】
次に、第２学習部４２が、マスクコーパス記憶部４１に記憶されたコーパスを用いて単語の意味表現を学習させて第２学習済みモデルを生成する（ステップ１０４）。この第２学習済みモデルは、第２学習済みモデル記憶部４３に記憶される。
【００６３】
次に、第２学習部４２が、第２学習済みモデル記憶部４３に記憶された第２学習済みモデルからユーザ既知語の意味表現を抜き出して第２出力情報として出力する（ステップ１０５）。この第２出力情報は、第２出力情報記憶部４５に記憶される。
【００６４】
次いで、質問生成装置１０では、第１出力情報記憶部２５に記憶された第１出力情報と第２出力情報記憶部４５に記憶された第２出力情報との出力差分を算出し、調査対象単語と関連付けて、出力差分情報として出力する（ステップ１０６）。この出力差分情報は、出力差分情報記憶部５２に記憶される。
【００６５】
その後、質問生成装置１０は、調査対象単語が終了したかどうかを判定する（ステップ１０７）。つまり、着目すべき調査対象単語がなくなったかどうかを判定する。
【００６６】
その結果、調査対象単語が終了していないと判定すれば、質問生成装置１０は、処理をステップ１０３へ戻す。そして、他の調査対象単語に着目し、ステップ１０３～１０６の処理を行う。
【００６７】
一方、調査対象単語が終了したと判定すれば、質問生成装置１０は、処理をステップ１０８へ進める。
【００６８】
そして、ランキング処理部５３が、調査対象単語を出力差分が大きい順に並べ替えて、質問順に並べられた質問単語として出力する（ステップ１０８）。この質問単語は、質問単語記憶部５４に記憶される。
【００６９】
［変形例］
上記実施の形態では言及しなかったが、システムは、ユーザから質問に対する回答が得られた時点で、新たなユーザ既知語を特定し、コーパス記憶部２１に記憶されたコーパスにこれを反映させてもよい。ここで、新たなユーザ既知語は、ユーザがタスク中でその単語の意味を知っているかを明示的にシステムに伝えることで、特定されるようにするとよい

。これにより、質問生成装置１０では、出力差分算出部５１が、この新たなユーザ既知語が反映されたコーパスを用いて新たに出力差分情報を生成することにより、ユーザ既知語を再度予測するようにしてよい。そして、ランキング処理部５３が、質問に用いる単語の順序をリアルタイムに更新してよい。この場合、出力差分算出部５１は、特定のユーザ既知語に代えて、質問に対するユーザの回答から把握される他のユーザ既知語を用いて、複数の差分を算出する手段の一例であり、ランキング処理部５３は、複数の差分に基づいて、複数の対象単語を用いた質問を再生成する手段の一例である。

【００７０】

［プロセッサ］

本実施の形態において、プロセッサとは広義的なプロセッサを指し、汎用的なプロセッサ（例えばＣＰＵ：Central Processing Unit等）や、専用のプロセッサ（例えばＧＰＵ：Graphics Processing Unit、ＡＳＩＣ：Application Specific Integrated Circuit、ＦＰＧＡ：Field Programmable Gate Array、プログラマブル論理デバイス等）を含むものである。

【００７１】

また、本実施の形態におけるプロセッサの動作は、１つのプロセッサによって成すのみでなく、物理的に離れた位置に存在する複数のプロセッサが協働して成すものであってもよい。また、プロセッサの各動作の順序は、本実施の形態において記載した順序のみに限定されるものではなく、変更してもよい。

【００７２】

［プログラム］

本実施の形態における質問生成装置１０が行う処理は、例えば、アプリケーションソフトウェア等のプログラムとして用意される。

【００７３】

即ち、本実施の形態を実現するプログラムは、コンピュータに、特定の例文集合を用いて単語の意味表現を学習させた第１のモデルから得られた特定のユーザ既知語の意味表現と、特定の例文集合の対象単語を除く部分を用いて単語の意味表現を学習させた第２のモデルから得られた特定のユーザ既知語の意味表現との差分を算出する機能と、差分に基づいて、対象単語がユーザ既知語である可能性に関する情報を出力する機能とを実現させるためのプログラムとして捉えられる。

【００７４】

尚、本実施の形態を実現するプログラムは、通信手段により提供することはもちろん、ＣＤ－ＲＯＭ等の記録媒体に格納して提供することも可能である。

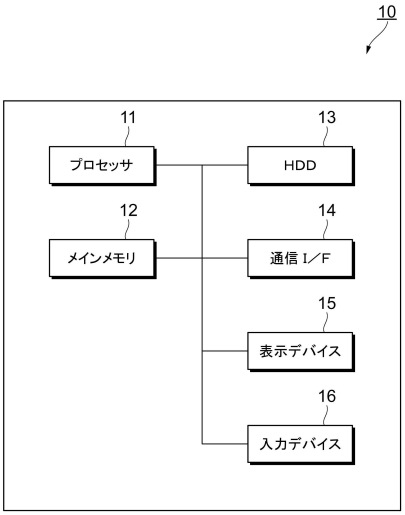
【符号の説明】

【００７５】

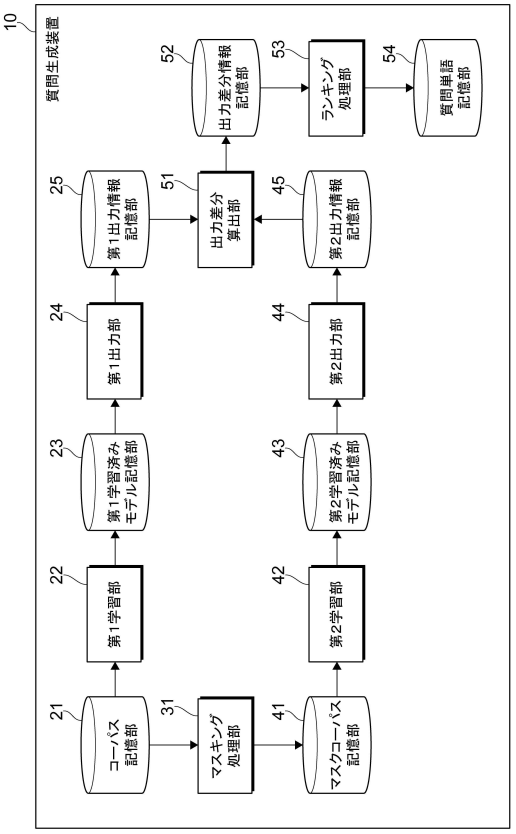
１０…質問生成装置、２１…コーパス記憶部、２２…第１学習部、２３…第１学習済みモデル記憶部、２４…第１出力部、２５…第１出力情報記憶部、３１…マスキング処理部、４１…マスクコーパス記憶部、４２…第２学習部、４３…第２学習済みモデル記憶部、４４…第２出力部、４５…第２出力情報記憶部、５１…出力差分算出部、５２…出力差分情報記憶部、５３…ランキング処理部、５４…質問単語記憶部

【図面】

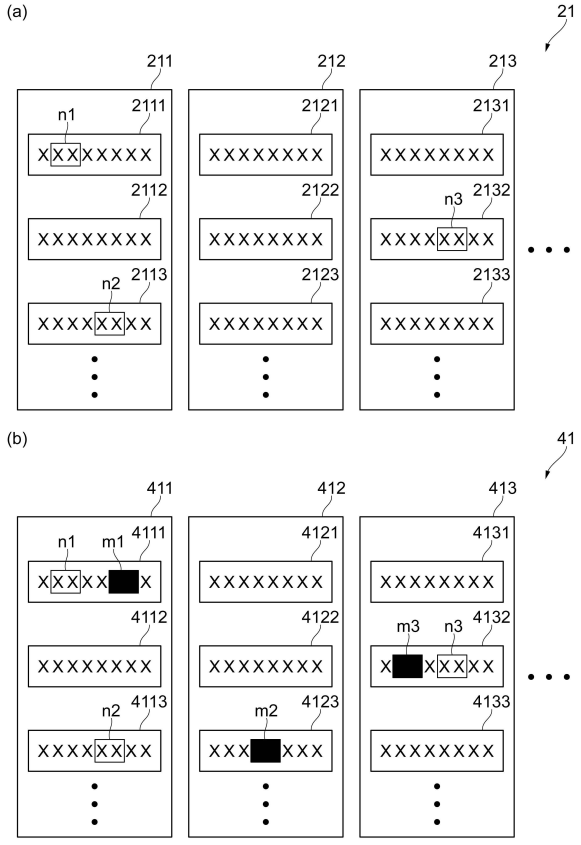
【図１】



【図２】



【図３】



【図４】

(a)

23

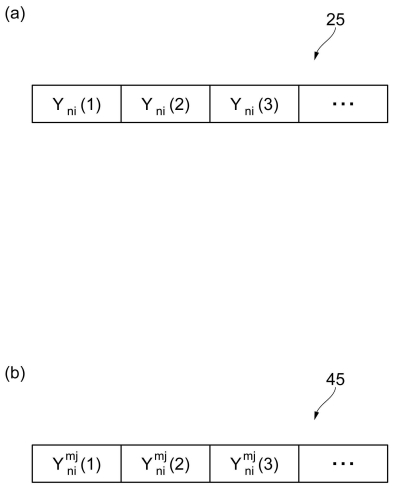
$Y_{v1}(1)$	$Y_{v1}(2)$	$Y_{v1}(3)$	…
$Y_{v2}(1)$	$Y_{v2}(2)$	$Y_{v2}(3)$	…
$Y_{v3}(1)$	$Y_{v3}(2)$	$Y_{v3}(3)$	…
…	…	…	…

(b)

43

$Y^{mj}_{v1}(1)$	$Y^{mj}_{v1}(2)$	$Y^{mj}_{v1}(3)$	…
$Y^{mj}_{v2}(1)$	$Y^{mj}_{v2}(2)$	$Y^{mj}_{v2}(3)$	…
$Y^{mj}_{v3}(1)$	$Y^{mj}_{v3}(2)$	$Y^{mj}_{v3}(3)$	…
…	…	…	…

【図５】



【図６】

52

調査対象単語	出力差分
m1	δ(ni,m1)
m2	δ(ni,m2)
m3	δ(ni,m3)
・・・	・・・

【図７】