

(19)日本国特許庁(JP)

(12)公開特許公報(A)

(11)公開番号

特開2023-168024

(P2023-168024A)

(43)公開日 令和5年11月24日(2023.11.24)

(51)国際特許分類

G 0 6 F 16/90 (2019.01)

F I

G 0 6 F

16/90

1 0 0

テーマコード (参考)

5 B 1 7 5

審査請求 未請求 請求項の数 9 O L (全26頁)

(21)出願番号 特願2022-79628(P2022-79628)

(22)出願日 令和4年5月13日(2022.5.13)

(71)出願人 318010018

キオクシア株式会社

東京都港区芝浦三丁目1番21号

(74)代理人 110003708

弁理士法人鈴榮特許総合事務所

(72)発明者 星 康人

東京都港区芝浦三丁目1番21号 キオ

クシア株式会社内

(72)発明者 宮下 大輔

東京都港区芝浦三丁目1番21号 キオ

クシア株式会社内

(72)発明者 出口 淳

東京都港区芝浦三丁目1番21号 キオ

クシア株式会社内

F ターム (参考) 5B175 EA01 HB03

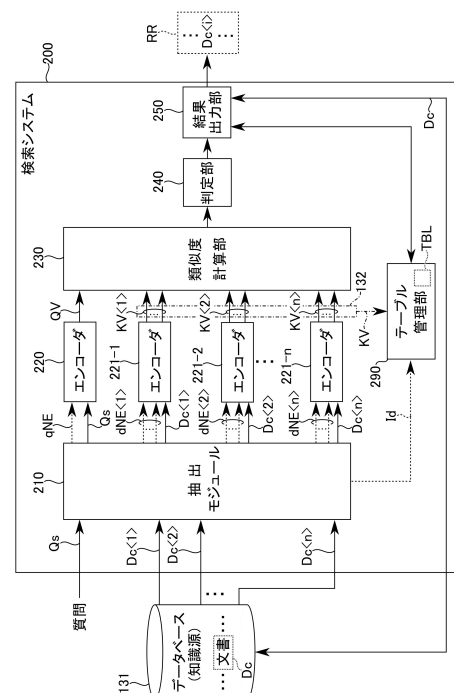
(54)【発明の名称】 情報処理装置及び情報処理方法

(57)【要約】 (修正有)

【課題】機械学習のタスクの精度を向上する情報処理装置及び情報処理方法を提供する。

【解決手段】情報処理装置において、検索システム200は、複数の文書に関する複数の第1のデータアイテムDc及び質問に関する第2のデータアイテムQsを受け、データアイテムDc、Qsを処理する抽出モジュールと、類似度計算部と、を含む。抽出モジュールは、第1のデータアイテムDcの夫々から複数の第1の固有表現dNEを抽出し、第2のデータアイテムQsから第2の固有表現qNEを抽出し、第1のデータアイテムの夫々に対応する複数の第1の固有表現の夫々に関する複数の第1のベクトルKVを生成し、第2のデータアイテムと第2の固有表現とに関する第2のベクトルQVを生成する。類似度計算部は、ベクトルKV、QV間の類似度を計算し、類似度の計算結果に基づいて、複数の第1のデータアイテムの中から回答に関する第3のデータアイテムRRを取得する。

【選択図】図2



【特許請求の範囲】

【請求項 1】

複数の文書に関する複数の第 1 のデータアイテム及び質問に関する第 2 のデータアイテムを受けるインターフェース回路と、

前記インターフェース回路が受けた前記複数の第 1 のデータアイテム及び前記第 2 のデータアイテムを処理するプロセッサと、

を備え、

前記プロセッサは、

前記複数の第 1 のデータアイテムのそれぞれから複数の第 1 の固有表現を抽出し、前記第 2 のデータアイテムから第 2 の固有表現を抽出し、

前記複数の第 1 のデータアイテムのそれぞれと対応する前記複数の第 1 の固有表現のそれぞれとに関する複数の第 1 のベクトルを、それぞれ生成し、

前記第 2 のデータアイテムと前記第 2 の固有表現とに関する第 2 のベクトルを生成し、

前記複数の第 1 のベクトルのそれぞれと前記第 2 のベクトルとの間の類似度を計算し、

前記類似度の計算結果に基づいて、前記複数の第 1 のデータアイテムの中から検索された回答に関する第 3 のデータアイテムを取得する、

情報処理装置。

【請求項 2】

前記プロセッサは、固有表現に関する第 1 の自然言語モデルの第 1 のニューラルネットワークに基づいて、前記複数の第 1 の固有表現及び前記第 2 の固有表現を抽出する、

請求項 1 に記載の情報処理装置。

【請求項 3】

前記プロセッサは、第 2 の自然言語モデルの第 2 のニューラルネットワークに基づいて、前記複数の第 1 のベクトル及び前記第 2 のベクトルを生成する、

請求項 1 に記載の情報処理装置。

【請求項 4】

前記プロセッサは、前記複数の第 1 のデータアイテムのうち 1 つの第 1 のデータアイテムから k 個 (k は、2 以上の整数) の前記第 1 の固有表現を抽出し、前記 1 つの第 1 のデータアイテム及び前記 k 個の前記第 1 の固有表現に基づいて k 個の前記第 1 のベクトルを生成する、

請求項 1 に記載の情報処理装置。

【請求項 5】

前記プロセッサは、複数の識別番号と前記複数の第 1 のベクトルとの対応関係を記憶し、

前記複数の識別番号のそれぞれは、前記複数の第 1 のデータアイテムのそれぞれに割り当てられている、

請求項 1 に記載の情報処理装置。

【請求項 6】

前記複数の第 1 のデータアイテムのそれぞれに複数の識別番号のそれぞれが割り当てられ、

前記複数の識別番号のそれぞれは、前記複数の第 1 のベクトルのうち対応する 1 つに関連付けられ、

前記複数の第 1 のベクトルのうち選択された第 1 のベクトルに対応する前記識別番号に基づいて、前記複数の第 1 のデータアイテムの中から前記選択されたベクトルに対応する第 1 のデータアイテムが、取得される、

請求項 1 に記載の情報処理装置。

【請求項 7】

前記複数の第 1 のデータアイテム及び前記複数の第 1 のベクトルを記憶するストレージを、

さらに具備する請求項 1 に記載の情報処理装置。

10

20

30

40

50

【請求項 8】

生成された前記複数の第 1 のベクトルを記憶するストレージをさらに具備し、
前記プロセッサは、前記類似度を計算する時、前記複数の第 1 のベクトルを前記ストレージから読み出す、
請求項 1 に記載の情報処理装置。

【請求項 9】

複数の文書に関する複数の第 1 のデータアイテムを受けると、
前記複数の第 1 のデータアイテムのそれぞれから複数の第 1 の固有表現を抽出することと、
前記複数の第 1 のデータアイテムのそれぞれと対応する前記複数の第 1 の固有表現のそれぞれとに基づいて、複数の第 1 のベクトルを生成することと、
質問に関する第 2 のデータアイテムを受けると、
前記第 2 のデータアイテムから第 2 の固有表現を抽出することと、
前記第 2 のデータアイテムと前記第 2 の固有表現とに基づいて、第 2 のベクトルを生成することと、
前記複数の第 1 のベクトルのそれぞれと前記第 2 のベクトルとの間の類似度を計算することと、
前記類似度の計算結果に基づいて、前記複数の第 1 のデータアイテムの中から検索された回答に関する第 3 のデータアイテムを取得することと、
を具備する情報処理方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明の実施形態は、情報処理装置及び情報処理方法に関する。

【背景技術】

【0002】

機械学習に関する方法、デバイス、及びシステムが、研究及び提案されている。例えば、機械学習の各種のタスクの精度の向上のために、様々な計算手法、処理手法、システムの構成、及びデバイスの構成が、研究及び提案されている。機械学習を用いた情報の検索において、入力データである質問に関する情報をデータベースの中から検索するタスクがある。このタスクにおいて、未知の質問に対する回答に関する情報の検索の精度が低い場合がある。そのため、機械学習を用いた情報の検索の精度の向上が求められている。

【先行技術文献】

【特許文献】

【0003】

【特許文献 1】特開 2020-140435 号

【特許文献 2】特開 2021-99803 号

【特許文献 3】特開 2019-164409 号

【発明の概要】

【発明が解決しようとする課題】

【0004】

機械学習のタスクの精度を向上する情報処理装置及び情報処理方法を提供する。

【課題を解決するための手段】

【0005】

本実施形態の情報処理装置は、複数の文書に関する複数の第 1 のデータアイテム及び質問に関する第 2 のデータアイテムを受け取るインターフェース回路と、前記インターフェース回路が受けた前記複数の第 1 のデータアイテム及び前記第 2 のデータアイテムを処理するプロセッサと、を含み、前記プロセッサは、前記複数の第 1 のデータアイテムのそれぞれから複数の第 1 の固有表現を抽出し、前記第 2 のデータアイテムから第 2 の固有表現を抽出し、前記複数の第 1 のデータアイテムのそれぞれと対応する前記複数の第 1 の固有表

現のそれぞれとに関する複数の第 1 のベクトルを、それぞれ生成し、前記第 2 のデータアイテムと前記第 2 の固有表現とに関する第 2 のベクトルを生成し、前記複数の第 1 のベクトルのそれぞれと前記第 2 のベクトルとの間の類似度を計算し、前記類似度の計算結果に基づいて、前記複数の第 1 のデータアイテムの中から検索された回答に関する第 3 のデータアイテムを取得する。

【図面の簡単な説明】

【0006】

【図 1】実施形態の情報処理装置の構成例を示すブロック図。

【図 2】実施形態の情報処理装置の検索システムの構成例を示すブロック図。

【図 3】実施形態の情報処理装置の検索システムの一部の構成例を示すブロック図。

10

【図 4】実施形態の情報処理装置の検索システムの他の一部の構成例を示すブロック図。

【図 5】実施形態の情報処理装置の検索システムを説明するための模式図。

【図 6】実施形態の情報処理方法を示すフローチャート。

【図 7】実施形態の情報処理方法の一部を説明するための模式図。

【図 8】実施形態の情報処理方法の他の一部を説明するための模式図。

【図 9】実施形態の情報処理方法を示すフローチャート。

【図 10】実施形態の情報処理方法の一部を説明するための模式図。

【図 11】実施形態の情報処理方法の他の一部を説明するための模式図。

【図 12】実施形態の情報処理方法の更に他の一部を説明するための模式図。

【図 13】実施形態の情報処理方法の更に他の一部を説明するための模式図。

20

【図 14】実施形態の情報処理装置の特性を示すグラフ。

【図 15】実施形態の情報処理装置の変形例を説明するための模式図。

【発明を実施するための形態】

【0007】

以下、図 1 乃至図 15 を参照しながら、実施の形態について詳細に説明する。以下の説明において、同一の機能及び構成を有する要素については、同一符号を付す。

また、以下の各実施形態において、末尾に区別化のための数字／英字を伴った参照符号を付された構成要素（例えば、回路、配線、各種の電圧及び信号など）が、相互に区別されなくとも良い場合、末尾の数字／英字が省略された記載（参照符号）が用いられる。

【0008】

30

（1）実施形態

図 1 乃至図 14 を参照して、実施形態の情報処理装置、及び、実施形態の情報処理方法について、説明する。尚、実施形態の情報処理方法は、実施形態の情報処理装置の制御方法を含み得る。

【0009】

（a）構成

図 1 乃至図 5 を参照して、実施形態の情報処理装置の構成及び機能について、説明する。

【0010】

図 1 は、本実施形態の情報処理装置 1 の構成例を説明するための模式図である。

40

【0011】

実施形態の情報処理装置 1 は、各種の情報処理を実行できる。情報処理装置 1 は、例えば、自然言語処理を実行可能な人工知能（AI）の機能を有する。

【0012】

情報処理装置 1 は、機械学習に基づいた各種の処理及びタスクを実行できる。例えば、情報処理装置 1 は、教師有り又は教師無しの学習データを用いた深層学習を実行可能に構成されている。

【0013】

情報処理装置 1 は、プロセッサ 11、メモリ 12、ストレージ 13、及びユーザーインターフェース 14 を含む。

50

【0014】

プロセッサ11は、情報処理装置1の各種の処理及びタスクの実行のための制御処理及び計算処理を、行う。プロセッサ11は、メモリ12、ストレージ13、及びユーザーインターフェース14を、制御できる。例えば、プロセッサ11は、各種の制御処理、各種の計算処理及び各種のタスクなどを実行する処理部200を含む。

プロセッサ11は、レジスタ、キャッシュ及びROM (Read only memory) などを含む記憶領域111を含む。

【0015】

また、プロセッサ11は、インターフェース部119を含む。インターフェース部119は、メモリ12、ストレージ13及びユーザーインターフェース14から各種のデータアイテムを受け取ることができる。インターフェース部119は、受け取ったデータアイテムを、記憶領域111及び（又は）処理部200に送ることができる。インターフェース部119は、例えば、処理部200によって実行された処理の結果のような、プロセッサ11の処理結果のデータアイテムを、メモリ12、ストレージ13及びユーザーインターフェース14に送ることができる。インターフェース部119は、これらの処理のための複数のデバイス及び回路などを含み得る。

【0016】

メモリ12は、各種の情報及び各種のデータアイテムを一時的に記憶する。

メモリ12は、RAM (Random access memory) 121及びROM 122を含む。

【0017】

RAM 121は、SRAM (Static RAM) 及びDRAM (Dynamic RAM) などを含む。RAM 121は、情報処理装置1に用いられる各種のデータアイテム及びソフトウェアなどを、一時的に記憶する。RAM 121は、情報処理装置1におけるワークメモリ及びバッファメモリとして機能する。RAM 121は、データアイテムの取得のために、プロセッサ11にアクセスされ得る。

【0018】

ROM 122は、情報処理装置1に用いられるオペレーティングシステム (OS)、ファームウェア、各種のソフトウェア及び各種のデータアイテムを実質的に不揮発に記憶する。ROM 122は、データアイテムの取得のために、プロセッサ11にアクセスされ得る。

【0019】

例えば、データアイテムは、処理の対象であるユーザーデータ、各種のシステム及びデバイスに用いられる設定データ、各種の処理に用いられるパラメータ、及びソフトウェアの一部などを含む。例えば、ソフトウェアは、実行プログラム、ファームウェア、アプリケーション及びオペレーティングシステム (OS) を含み得る。データアイテム及び（又は）ソフトウェアは、各種のシステム及びデバイスに用いられる情報に相当し得る。

【0020】

ストレージ13は、各種の情報及び各種のデータアイテムを実質的に不揮発に記憶できる。ストレージ13は、複数のデータベース131, 132を含む。

【0021】

ストレージ13は、例えば、SSD (Solid state drive)、UFS (Universal Flash Storage) デバイス又はHDD (Hard disc drive) である。ストレージ13がSSD又はUFSデバイスである場合、ストレージ13は、コントローラ及び不揮発性半導体メモリデバイスを含む。SSDに用いられる不揮発性半導体メモリデバイスは、NAND型フラッシュメモリである。不揮発性半導体メモリデバイスが、NAND型フラッシュメモリである場合、不揮発性半導体メモリデバイスは、Toggle DDRインターフェース規格又はONFi (Open NAND Flash interface) 規格に基づいて、コントローラと通信する。コントローラは、読み出しシーケンス、書き込みシーケンス、消去シーケンス及びガベージコレクションのような各種の動作を、不揮発性半導体メモリデバイスに命令する。不揮発性半導体メモリデバイスは、コントローラからの命令に基づいて、各

10

20

30

40

50

種の動作を実行する。

【0022】

尚、本実施形態の情報処理装置1において、ストレージ13は、プロセッサ11と直接的又は間接的に通信が可能であれば、1つの筐体（図示せず）内に設けられていてもよいし、互いに異なる筐体内に設けられていてもよい。プロセッサ11及びストレージ13は、直接的又は間接的に互いに通信が可能であれば、同じ国又は地域に設置されていてもよいし、互いに異なる国又は地域に設置されていてもよい。

【0023】

ユーザーインターフェース14は、情報処理装置1のユーザーとプロセッサ11との間の通信のための複数のデバイス及び回路などを含み得る。ユーザーインターフェース14は、入力デバイス及び出力デバイスを含む。入力デバイスは、例えば、タッチパネル、キーボード、操作ボタン及び（又は）マイクなどを含む。出力デバイスは、例えば、ディスプレイ、プリンタ及び（又は）スピーカーなどを含む。

【0024】

ユーザーインターフェース14は、入力デバイスを介してユーザーからの各種の処理及び（又は）タスクの実行の要求をプロセッサ11に供給する。ユーザーインターフェース14は、出力デバイスを介して各種の処理及び（又は）タスクの実行の結果をユーザーに提供する。ユーザーインターフェース14は、各種の処理及び（又は）タスクの実行の途中経過をユーザーに提供することもできる。

【0025】

ユーザーインターフェース14は、各種のデータアイテムを、プロセッサ11、メモリ12及びストレージ13を送ることができる。ユーザーインターフェース14は、各種のデータアイテムを、プロセッサ11、メモリ12及びストレージ13から受けることができる。

【0026】

情報処理装置1は、無線又は有線のネットワークを介して、他のデバイス9と通信してもよい。例えば、処理及び（又は）タスクの実行の要求が、デバイス9から情報処理装置1に供給されてもよい。処理及び（又は）タスクの実行の結果が、情報処理装置1からデバイス9に供給されてもよい。デバイス9は、コンピュータデバイス又は携帯デバイスである。コンピュータデバイスの一例は、パーソナルコンピュータ又はサーバコンピュータである。携帯デバイスの一例は、スマートフォン、フィーチャーフォン、又はタブレットデバイスである。

【0027】

本実施形態の情報処理装置1は、深層学習を用いた文書検索を実行するシステムである処理部200を含む。以下において、深層学習を用いた文書検索を実行するシステムである処理部200は、検索システム200とよばれる。

【0028】

検索システム200は、プロセッサ11上において、ソフトウェア、ハードウェア、又は、ソフトウェアとハードウェアとの組み合わせによって実現され得る。

これによって、本実施形態の情報処理装置1において、プロセッサ11は、文書の検索のための検索タスクの各種の処理及び動作を実行する。以下において、検索タスクは、検索処理ともよばれる。

【0029】

図2は、本実施形態の情報処理装置1における、検索システム200を説明するための模式的なブロック図である。

【0030】

図2に示されるように、本実施形態の情報処理装置1において、検索システム200は、深層学習に基づく文書の検索タスクを実行する。検索システム200は、検索タスクによって、質問Qsに対して、知識源としてのデータベース131に格納された複数の文書Dcの中から回答となる1つ以上の文書Dcを、選択及び提示する。

【0031】

質問Q_sは、自然言語の1つ以上の文を含むデータアイテムである。質問Q_sは、質問データアイテムともよばれる。例えば、質問Q_sは、ユーザーインターフェース14を介して、情報処理装置1の外部（例えば、ユーザー）から検索システム200に供給される。

【0032】

データベース131に格納された複数の文書D_cのそれぞれは、自然言語の1つ以上の文を含むデータアイテムである。文書D_cは、文書データアイテムともよばれる。例えば、複数の文書D_cは、ユーザーインターフェース14を介して、情報処理装置1の外部からストレージ13のデータベース131に供給される。

10

【0033】

本実施形態において、情報処理装置1及び検索システム200に用いられる自然言語は、日本語、英語、中国語、ドイツ語、フランス語、ロシア語、スペイン語、及びイタリア語などの複数の自然言語の中から選択されるいずれかの自然言語である。質問Q_sに用いられる自然言語は、文書D_cに用いられる自然言語と同じであることが、好ましい。但し、質問Q_sに用いられる自然言語は、文書D_cに用いられる自然言語と異なってもよい。

【0034】

図2の検索システム200は、抽出モジュール210、エンコーダ220、221（221-1～221-n）、類似度計算部230、判定部240、及び結果出力部250などを含む。例えば、検索システム200の各構成要素210、220、221、230、240、250は、プロセッサ11によって制御され得る。

20

【0035】

検索システム200は、質問Q_s及び複数の文書D_cを、受ける。

【0036】

質問Q_sは、ユーザーインターフェース14を介して、ユーザーから検索システム200に供給される。質問Q_sは、プロセッサ11の記憶領域111又はメモリ12に記憶される。質問Q_sは、他のデバイス9から、ネットワーク（図示せず）を介して、検索システム200に供給されてもよい。

【0037】

質問Q_sは、人物、組織、地域、事件、物品及び日時など或る事物に関して問う文を含む。質問Q_sは、自然言語によって記述されている。

30

【0038】

文書D_cは、ストレージ13のデータベース131から検索システム200に供給される。文書D_cは、人物、地域、事件及び物品などの或る事物に関する情報を示す文を含む。文書D_cは、自然言語によって記述されている。文書D_cは、1つ以上の文を含む。文書D_cは、複数の単語を含んで構成される。例えば、文書D_cは、十数から百数十の単語から構成される。文書D_cは、テキスト又はパッセージともよばれる。

【0039】

データベース131は、文書集合である。データベース131は、検索システム200における検索タスクのための知識源及び情報源として機能する。

40

【0040】

抽出モジュール210は、検索システム200に供給された質問Q_s及び文書D_cを受け、抽出モジュール210は、質問Q_s及び文書D_cから各種の情報を抽出できる。

【0041】

本実施形態において、抽出モジュール210は、或る質問Q_sに含まれる1つ以上の固有表現qNEを抽出できる。本実施形態において、抽出モジュール210は、或る文書D_cに含まれる1つ以上の固有表現dNEを、抽出できる。

固有表現qNE、dNEは、人名、組織名、地名、事件名、物品名、及び（又は）日時などである。

【0042】

50

抽出モジュール 210 は、抽出部又は固有表現抽出モジュールともよばれる。

【0043】

例えば、抽出モジュール 210 は、自然言語に関するモデルに基づくニューラルネットワークから構成された機構を含む。

抽出モジュール 210 の構成及び機能の詳細は、後述される。

【0044】

エンコーダ 220, 221 は、供給された自然言語のデータアイテム Q_s , D_c 及び固有表現 q_{NE} , d_{NE} から、ベクトル Q_V , K_V を生成する。

【0045】

エンコーダ 220 は、或る質問 Q_s の文及び質問 Q_s から抽出された固有表現 q_{NE} に基づいて、質問 Q_s に関するベクトル Q_V を生成する。以下において、質問 Q_s 及び固有表現 q_{NE} から得られたベクトル Q_V は、クエリベクトル Q_V とよばれる。尚、クエリベクトル Q_V は、クエリとよばれてもよい。

【0046】

クエリベクトル Q_V は、自然言語の質問 Q_s とその質問 Q_s に含まれる固有表現 q_{NE} に関する特徴量ベクトルのデータアイテムである。

【0047】

エンコーダ 221 は、或る文書 D_c の文及び文書 D_c から抽出された固有表現 d_{NE} に基づいて、文書 D_c に関するベクトル K_V を生成する。以下において、文書 D_c 及び固有表現 d_{NE} から得られたベクトル K_V は、キーベクトル K_V とよばれる。尚、キーベクトル K_V は、キーとよばれてもよい。

【0048】

キーベクトル K_V は、自然言語の文書 D_c とその文書 D_c に含まれる固有表現 d_{NE} に関する特徴量ベクトルのデータアイテムである。

例えば、キーベクトル K_V は、ストレージ 13 のデータベース 132 に記憶されている。

【0049】

例えば、エンコーダ 220, 221 のそれぞれは、自然言語に関するモデルに基づくニューラルネットワークから構成された機構を含む。

これによって、エンコーダ 220 は、自然言語の質問 Q_s 及び固有表現 q_{NE} を、ベクトル Q_V に変換できる。エンコーダ 221 は、自然言語の文書 D_c 及び固有表現 d_{NE} を、ベクトル K_V に変換できる。

エンコーダ 220, 221 の構成及び機能の詳細は、後述される。

【0050】

本実施形態において、例えば、キーベクトル K_V は、文書 D_c の特徴を示すデータアイテムとして用いられる。本実施形態において、例えば、クエリベクトル Q_V は、キーベクトル K_V 及び文書 D_c の検索に用いられるデータアイテムとして用いられる。

【0051】

類似度計算部 230 は、或るデータと別のデータとの間の類似度を計算する。本実施形態において、類似度計算部 230 は、クエリベクトル Q_V とキーベクトル K_V との間の類似度を計算する。

【0052】

例えば、類似度は、2つのベクトル Q_V , K_V 間のコサイン類似度に基づいて、計算される。尚、類似度は、2つのベクトル Q_V , K_V 間の内積、又は、2つのベクトル Q_V , K_V 間の距離などに基づいて、計算されてもよい。類似度を計算するための距離は、例えば、ユークリッド距離、マンハッタン距離及びミンコフスキー距離などのうちいずれか1つを用いて、得られる。

【0053】

判定部 240 は、プロセッサ 11 によって実行された各種の処理に対する判定を、実行できる。本実施形態において、判定部 240 は、類似度計算部 230 の計算結果に基づい

10

20

30

40

50

て、クエリベクトル QV とキーベクトル KV とが類似しているか否かを判定する。

【0054】

判定部 240 は、クエリベクトル QV 及びキーベクトル KV に関して計算された類似度の計算結果の値が或る閾値以上である場合、クエリベクトル QV がキーベクトル KV に類似していると判定する。判定部 240 は、クエリベクトル QV 及びキーベクトル KV に関して計算された類似度の計算結果の値が或る閾値未満である場合、クエリベクトル QV がキーベクトル KV に類似していないと判定する。以下において、類似度の計算結果の値は、スコアともよばれる。

【0055】

尚、複数の閾値が、2つのベクトル QV , KV 間の類似度を評価するために設定された場合、2つのベクトル QV , KV 間の類似度は、段階的に評価され得る。この結果として、回答となり得る複数の文書 Dc は、類似度のスコアに応じて、複数のレベルに分類され得る。

【0056】

結果出力部 250 は、判定部 240 の処理の結果に基づいて、検索タスクにおける質問 Qs に対する検索結果 RR として、或るスコア以上の類似度を有する1つ以上のキーベクトル $KV < i >$ のそれぞれに対応する1つ以上の文書 $Dc < i >$ を、出力する。 i は、1以上の整数である。

【0057】

例えば、結果出力部 250 は、各キーベクトル KV に対応する文書 Dc を、データベース 131 を含むストレージ 13 から読み出す。結果出力部 250 は、読み出した文書 Dc を、ユーザーインターフェース 14 を介して、ユーザーに示す。

【0058】

結果出力部 250 は、データベース 131 の文書 Dc を読み出す時、テーブル TBL に基づいて、検索結果のキーベクトル KV に対応する文書 Dc を検知する。

【0059】

テーブル TBL は、キーベクトル KV と文書 Dc との対応関係に関する情報を含む。例えば、テーブル TBL は、テーブル管理部 290 によって、生成される。テーブル TBL は、テーブル管理部 290 に記憶される。

【0060】

テーブル管理部 290 は、抽出モジュール 210 から提供された各文書 Dc に関する情報及びエンコーダ 221 から提供されたキーベクトル KV に関する情報に基づいて、テーブル TBL を生成できる。

例えば、テーブル TBL は、管理テーブル TBL ともよばれる。

【0061】

例えば、テーブル管理部 290 は、プロセッサ 11、メモリ 12、及びストレージ 13 のうちいずれか1つに設けられ得る。

【0062】

このように、検索システム 200 は、質問 Qs に対して高い類似度を有する文書 Dc を、データベース 131 の中から検索する。

検索システム 200 は、検索タスクの検索結果 RR を、ユーザーに示す。この結果として、検索結果 RR が、質問 Qs に対する回答として、ユーザーに提供される。

【0063】

尚、質問 Qs は、ネットワークを介して他のデバイス 9 から情報処理装置 1 に供給されてもよい。質問 Qs に対する検索結果 RR は、ネットワークを介して情報処理装置 1 から他のデバイス 9 に供給されてもよい。

【0064】

抽出モジュール 210 及びエンコーダ 220 , 221 は、ソフトウェア又はファームウェアとして、プロセッサ 11 に、提供される。抽出モジュール 210 及びエンコーダ 220 は、例えば、Python のような或るプログラム言語によって形成されたコンピュータ

10

20

30

40

50

ログラムとして、プロセッサ 11 の記憶領域 111 に記憶されている。

【0065】

抽出モジュール 210 及びエンコーダ 220, 221 のソフトウェアは、メモリ 12 に記憶されてもよいし、ストレージ 13 に記憶されてもよい。この場合、それらのソフトウェアが、後述される抽出モジュール 210 及びエンコーダ 220, 221 を用いた処理の実行時に、メモリ 12 からプロセッサ 11 の記憶領域 111 に、又は、ストレージ 13 からプロセッサ 11 の記憶領域 111 に読み出される。

【0066】

尚、抽出モジュール 210 及びエンコーダ 220, 221 のソフトウェアは、抽出モジュール 210 及びエンコーダ 220, 221 を用いた後述の処理の実行時に、RAM 131 に記憶され、それらのソフトウェアが、プロセッサ 11 によって RAM 131 上で実行されてもよい。

抽出モジュール 210 及びエンコーダ 220, 221 は、ハードウェアとして、プロセッサ 11 の内部又はプロセッサ 11 の外部に設けられてもよい。

【0067】

＜抽出モジュール 210＞

図 3 は、本実施形態の情報処理装置 1 における、検索システム 200 の抽出モジュール 210 の構成及び機能を説明するためのブロック図である。

【0068】

抽出モジュール 210 は、質問受信部 211、複数の文書受信部 212 (212-1 ~ 212-n)、固有表現抽出部 213, 214 (214-1 ~ 214-n)、及び、識別番号抽出部 219 (219-1 ~ 219-n) などを含む。n は、1 以上の整数である。

【0069】

質問受信部 211 は、検索システム 200 に供給された質問 Qs を受ける。

【0070】

複数の文書受信部 212-1 ~ 212-n のそれぞれは、データベース 131 から供給された文書 Dc を、受ける。

【0071】

例えば、文書 Dc は、識別番号 Id を含む。識別番号 Id は、文書 Dc に付加された情報である。識別番号 Id と文書 Dc とを含むデータアイテムは、文書ファイル FL (FL<1> ~ FL<n>) ともよばれる。

【0072】

識別番号 Id は、データベース 131 において複数の文書 Dc を管理及び識別するための番号を示すデータアイテムである。各識別番号 Id は、複数の文書 Dc のうち対応する 1 つの文書 Dc に割り当てられている。識別番号 Id は、文書 Dc ごとに固有の値を有する。識別番号 Id は、文書ファイル FL のファイル名でもよいし、文書ファイル FL のメタデータに含まれる値でもよい。識別番号 Id は、文書 Dc に組み込まれた値でもよい。

【0073】

固有表現抽出部 213, 214 のそれぞれは、固有表現抽出モデル M1 に基づくニューラルネットワークを含む。

固有表現抽出部 213, 214 のそれぞれは、固有表現抽出モデル M1 に基づいて、固有表現 qNE, dNE を抽出する。固有表現抽出モデル M1 は、自然言語で記述された文中の語句が固有表現であるか否かを判別することが可能なように構築された、自然言語に関するニューラルネットワークモデルである。

【0074】

固有表現抽出部 213 は、固有表現抽出モデル M1 に基づいて、質問受信部 211 から質問 Qs を受ける。

固有表現抽出部 213 は、固有表現抽出モデル M1 に基づいて、質問 Qs を構成する文の中から、1 つの固有表現 qNE を抽出する。

【0075】

10

20

30

40

50

例えば、固有表現抽出部 213 は、質問 Qs を構成する文を、トークン化する。これによって、質問 Qs は、複数のトークンに分割される。

固有表現抽出部 213 は、固有表現抽出モデル M1 に基づいて、質問 Qs におけるトークンの位置、トークンの品詞、及び複数のトークン間の関係性などに基づいて、各トークンが固有表現であるか否かを判別する。

【0076】

この結果として、固有表現であると判別されたトークンが、質問 Qs における固有表現 qNE として、質問 Qs から抽出される。

固有表現抽出部 213 は、質問 Qs から抽出された固有表現 qNE を、エンコーダ 220 に送る。

【0077】

複数の固有表現抽出部 214-1 ~ 214-n のそれぞれは、複数の文書受信部 212-1 ~ 212-n のうち対応する 1 つから、文書 Dc を受ける。

各固有表現抽出部 214 は、文書 Dc を構成する 1 つ以上の文の中から、1 つ以上の固有表現 dNE ($dNE < 1 > - 1$, $dNE < 1 > - j$, ..., $dNE < n > - 1$, $dNE < n > - k$) を抽出する。j 及び k は、1 以上の整数である。

【0078】

例えば、各固有表現抽出部 214 は、文書 Dc に含まれる文を、トークン化する。これによって、文書 Dc は、複数のトークンに分割される。

固有表現抽出部 214 は、固有表現抽出モデル M1 に基づいて、文書 Dc の文におけるトークンの位置、トークンの品詞、及び複数のトークン間の関係性などに基づいて、各トークンが固有表現であるか否かを判別する。

【0079】

この結果として、固有表現であると判別されたトークンが、文書 Dc における固有表現 dNE として、文書 Dc から抽出される。

各固有表現抽出部 214 は、文書 Dc から抽出された 1 つ以上の固有表現 dNE を、複数のエンコーダ 221 のうち対応する 1 つに送る。

【0080】

複数の識別番号抽出部 219-1 ~ 219-n のそれぞれは、複数の文書受信部 212-1 ~ 212-n のうち対応する 1 つから、識別番号 Id が付加された文書 Dc (文書ファイル FL) を受ける。

各識別番号抽出部 219-1, 219-2, ..., 219-n のそれぞれは、識別番号 Id を、抽出する。識別番号抽出部 219 は、抽出した識別番号 Id を、テーブル管理部 290 に送る。抽出された識別番号 Id は、対応する文書 Dc 及び対応する文書 Dc から生成されるキーベクトル KV と紐づけされている。

【0081】

例えば、文書受信部 212 が、文書 Dc から識別番号 Id を分離することによって、識別番号 Id を抽出してもよい。

【0082】

尚、文書 Dc は、対応する識別番号 Id を含まない場合もある。この場合において、識別番号抽出部 219 が、文書 Dc に対応する識別番号 Id を、生成してもよい。

【0083】

質問受信部 211 は、質問 Qs を、固有表現抽出部 213 だけでなくエンコーダ 220 に送る。エンコーダ 220 は、固有表現 qNE とは別途に、質問 Qs を受ける。

【0084】

各文書受信部 212 は、文書 Dc を、対応する固有表現抽出部 214 だけでなく複数のエンコーダ 221 のうち対応する 1 つに送る。エンコーダ 221 は、固有表現 dNE とは別途に、文書 Dc を受ける。

【0085】

このように、抽出モジュール 210 は、質問 Qs 及び文書 Dc の文中から固有表現 qN

10

20

30

40

50

E, dNEを抽出する。

抽出モジュール210は、質問Qsと共に、抽出された固有表現qNEを、後段のエンコーダ220に送る。

抽出モジュール210は、文書Dcと共に、抽出された固有表現dNEを、後段のエンコーダ221に送る。

【0086】

<エンコーダ220, 221>

図4は、本実施形態の情報処理装置1における、検索システム200のエンコーダ220, 221の構成及び機能を説明するための図である。

【0087】

エンコーダ220, 221は、自然言語の深層学習に関する各種の計算処理を行うように構成されたニューラルネットワークを含む。

【0088】

図4の(a)は、本実施形態における、質問Qsに対するエンコーダ220を説明するための図である。

【0089】

エンコーダ220は、学習済み言語モデルM2のニューラルネットワークに基づいて、供給された自然言語のベクトル変換処理を行う。尚、学習済み言語モデルM2は、各種のデータセットを用いて自然言語処理に関する機械学習（深層学習）が既になされたニューラルネットワークモデルである。

【0090】

エンコーダ220は、ニューラルネットワークの構成に応じて複数のレイヤLY1a, LY2aを含む。例えば、レイヤLY1aは、埋め込みレイヤである。レイヤLY2aは、隠れレイヤである。尚、レイヤLY1aは、アテンションレイヤでもよい。レイヤLY2aは、フィードフォワードレイヤでもよい。

エンコーダ220は、質問Qs、及び、抽出モジュール210によって抽出された固有表現qNEを受ける。

【0091】

エンコーダ220は、質問Qsの文を複数のトークンTkに分割する。

エンコーダ220は、固有表現qNE及びトークン化された質問Qsに対して、レイヤLY1a, LY2aの機能に応じた各種の処理を行う。

【0092】

例えば、埋め込みレイヤを含むレイヤLY1aは、トークン化された質問Qs及び固有表現qNEに対して、各トークンの埋め込みのための各種の処理を行う。

例えば、隠れレイヤを含むレイヤLY2aは、レイヤLY1aからの出力に対して、積和演算のような各種の計算処理を行う。

【0093】

エンコーダ220は、レイヤLY1a, LY2aによる各種の処理によって、質問Qs及び固有表現qNEに基づくクエリベクトルQVを、生成及び出力する。クエリベクトルQVは、複数の成分Cmを含む。各成分Cmは、 $p \times q$ の行列式で示され得る。

【0094】

p及びqは、1以上の整数である。尚、本実施形態において、各成分Cmは、簡略化のためにハッチングが付された矩形パターンで模式的に図示されているが、 $p \times q$ で配列された複数の数値の配列である。

【0095】

図4の(b)及び(c)は、本実施形態における、文書Dcに対するエンコーダ221を説明するための図である。

【0096】

図4の(b)に示されるように、エンコーダ221は、エンコーダ220と実質的に同じ構成及び機能を有する。

10

20

30

40

50

エンコーダ 221 は、学習済み言語モデル M2 のニューラルネットワークに基づいて、供給された自然言語のベクトル変換処理を行う。

エンコーダ 221 は、複数のレイヤ LY1b, LY2b を含む。

【0097】

エンコーダ 221 は、文書 Dc、及び、抽出モジュール 210 によってその文書 Dc から抽出された複数の固有表現 dNEa, dNEb のうち固有表現 dNEa を受ける。

【0098】

エンコーダ 221 は、文書 Dc の文を複数のトークン Tk に分割する。

エンコーダ 221 は、固有表現 dNEa 及びトークン化された文書 Dc に対して、各レイヤ LY1b, LY2b の機能に応じた各種の処理を行う。

【0099】

エンコーダ 221 は、レイヤ LY1b, LY2b による各種の処理によって、文書 Dc 及び固有表現 dNEa に基づくキーベクトル KV a を、生成及び出力する。キーベクトル KV a は、 $p \times q$ の成分 Cm を含む。キーベクトル KV a の成分 Cm の数は、クエリベクトル QV の成分 Cm の数と同じである。

【0100】

上述のように、1つの文書 Dc から抽出された複数の固有表現 dNE (dNEa, dNEb) のうちの1つが、1つのエンコーダ 221 に、供給される。

エンコーダ 221 は、固有表現 dNE 毎に、文書 Dc と固有表現 dNE とのキーベクトル KV を生成する。

【0101】

図 4 の (c) に示されるように、文書 Dc から抽出された固有表現 dNEa とは別の固有表現 dNEb が、文書 Dc と共に、別のエンコーダ 221 に供給される。

【0102】

このエンコーダ 221 は、文書 Dc 及び固有表現 dNEb に基づくキーベクトル KV b を、生成及び出力する。

キーベクトル KV b の成分 Cm の数は、キーベクトル KV a の成分 Cm の数と同じである。キーベクトル KV b の複数の成分 Cm は、キーベクトル KV a の複数の成分 Cm と異なる。

【0103】

図 4 の (b) 及び (c) のように、1つの文書 Dc が 2つの固有表現 dNEa, dNEb を含んでいる場合、固有表現 dNEa, dNEb 毎に、2つのキーベクトル KV a, KV b が、生成される。

【0104】

複数のキーベクトル KV は、ストレージ 13 のデータベース 132 に記憶される。

【0105】

生成された複数のキーベクトル KV は、潜在表現空間 LS を形成する。潜在表現空間 LS は、学習済み言語モデル M2 のような、共通のモデルのニューラルネットワークに基づいて生成されたベクトルの集合である。尚、潜在表現空間 LS は、潜在空間 LS ともよばれる。

【0106】

本実施形態における質問 Qs と固有表現 qNE とから生成されるクエリベクトル QV は、或る固有表現 qNE を含む質問 Qs の文脈がその固有表現 qNE に反映されることによって生成されるベクトルである、とも換言できる。

本実施形態における文書 Dc と固有表現 dNE とから生成されるキーベクトル KV は、或る固有表現 dNE を含む文書 Dc の文脈がその固有表現 dNE に反映されることによって生成されるベクトルである、とも換言できる。

【0107】

以上のように、本実施形態において、検索システム 200 に用いられる複数のベクトル QV, KV が、形成される。

10

20

30

40

50

【0108】

尚、上述の例において、文書 D c 毎に、対応する 1 つのエンコーダ 221 が設けられた構成が、示されている。

但し、複数の文書 D c が、1 つのエンコーダ 221 に対応付けられてもよい。この場合において、1 つのエンコーダ 221 が、順次供給される複数の文書 D c のそれぞれに対して、各文書 D c と固有表現 d N E とを用いて、複数のキーベクトル K V を順次生成する。

【0109】

<テーブル T B L>

図 5 は、本実施形態の情報処理装置 1 における、検索システム 200 が文書 D c 及びキーベクトル K V を管理するためのテーブル T B L を説明するための図である。

10

【0110】

例えば、テーブル管理部 290 は、抽出モジュール 210 からの複数の文書 D c の複数の識別番号 I d に関する情報と、複数のエンコーダ 221 からの複数のキーベクトル K V に関する情報と、を受ける。

【0111】

テーブル管理部 290 は、複数の識別番号 I d に関する情報及び複数のキーベクトル K V に関する情報に基づいて、図 5 のようなテーブル T B L を生成及び格納する。

【0112】

図 5 は、本実施形態の情報処理装置 1 における、テーブル T B L の一例を示している。

【0113】

図 5 に示されるように、或る文書 D c (D c < 1 > , D c < 2 > , D c < 3 > , . . . , D c < n - 1 > , D c < n >) 及びその文書 D c の識別番号 I d (I d < 1 > , I d < 2 > , I d < 3 > , . . . , I d < n - 1 > , I d < n >) に基づいて、その文書 D c から生成された 1 つ以上のキーベクトル K V (K V < 1 > , K V < 2 > , K V < 3 > , . . . , K V < n - 1 > , K V < n >) は、その文書 D c の 1 つの識別番号 I d に関連付けられる。

20

【0114】

上述のように、各識別番号 I d < 1 > , I d < 2 > , I d < 3 > , . . . , I d < n - 1 > , I d < n > は、各文書 D c < 1 > , D c < 2 > , D c < 3 > , . . . , D c < n - 1 > , D c < n > に、1 対 1 で関連付けられている。

30

【0115】

図 5 のテーブル T B L において、1 つの識別番号 I d に、1 つの文書 D c から生成された 1 つ以上のキーベクトル K V が関連付けられる。

例えば、< 1 > の識別番号 I d に、2 つのキーベクトル K V < 1 > - 1 , K V < 1 > - 2 が関連付けられている。例えば、< 2 > の識別番号 I d に、2 つのキーベクトル K V < 2 > - 1 , K V < 2 > - 2 が関連付けられている。例えば、< 3 > の識別番号 I d に、m 個のキーベクトル K V < 3 > - 1 , K V < 3 > - 2 , . . . , K V < 3 > - m が関連付けられている。m は、1 以上の整数である。

例えば、< n - 1 > の識別番号 I d に、1 つのキーベクトル K V < n - 1 > が関連付けられている。例えば、< n > の識別番号 I d に、2 つのキーベクトル K V < n > - 1 , K V < n > - 2 が関連付けられている。

40

【0116】

それゆえ、テーブル管理部 290 に保存されたテーブル T B L に基づいて、対応する識別番号 I d が、選択された 1 つのキーベクトル K V から検知できる。

【0117】

したがって、1 つの文書 D c から複数のキーベクトル K V が生成されたとしても、対応する文書 D c が、或るキーベクトル K V に関連付けられた識別番号 I d に基づいて、データベース 131 の中から読み出され得る。

【0118】

このように、本実施形態において、検索システム 200 は、テーブル T B L の参照によ

50

って、キーベクトル KV に対応する特定の文書 Dc を、取得することができる。

【0119】

本実施形態のように、キーベクトル KV に基づく文書 Dc の検索及び取得が、識別番号 Id を用いて実行された場合、検索システム 200 の構成を、簡素化できる。例えば、本実施形態において、検索システム 200 は、キーベクトル KV を文に変換するためのデコーダを、含まない。

【0120】

以上のように、本実施形態の情報処理装置 1 における検索システム 200 は、上述の構成及び機能を含む。

【0121】

本実施形態において、生成されたクエリベクトル QV は、質問 Qs の文の文脈の特徴量ベクトルに、質問 Qs に含まれる固有表現 qNe の特徴量ベクトルが反映された値を有する。

これと同様に、生成されたキーベクトル KVa , KVb のそれぞれは、文書 Dc の文の文脈の特徴量ベクトルに、文書 Dc に含まれる各固有表現 $dNEa$, $dNEb$ の特徴量ベクトルが反映された値を有する。

【0122】

互いに異なる固有表現 $dNEa$, $dNEb$ の特徴がそれぞれ反映された 2 つのキーベクトル KVa , KVb は、1 つの文書 Dc に基づく特徴量ベクトルであっても、固有表現 $dNEa$, $dNEb$ 毎に異なる成分 Cm を含み得る。

【0123】

例えば、本実施形態において、文書 Dc の文脈が質問 Qs の文脈に関連していても、文書 Dc に含まれる固有表現 dNE が質問 Qs に含まれる固有表現 qNE と類似していなければ、固有表現 dNE の特徴が反映されたキーベクトル KV は、クエリベクトル QV に対して低い類似度を、有する。この結果として、質問 Qs の固有表現 qNE に対して低い類似度の固有表現 dNE を有する文書 Dc は、質問 Qs に対する回答から除外される。

【0124】

これに対して、文書 Dc の文脈が質問 Qs の文脈に関連し、且つ、文書 Dc に含まれる固有表現 dNE が質問 Qs に含まれる固有表現 qNE と類似していれば、固有表現 dNE の特徴が反映されたキーベクトル KV は、クエリベクトル QV に対して高い類似度を、有する。この結果として、質問 Qs の固有表現 qNE に対して高い類似度の固有表現 dNE を有する文書 Dc は、回答として選択される。

【0125】

また、本実施形態の情報処理装置 1 によれば、1 つの文書 Dc に含まれる固有表現 dNE の数に応じて、1 つの文書 Dc から複数のキーベクトル KV が、生成され得る。

これによって、1 つの文書 Dc から 1 つのキーベクトル KV が生成される場合に比較して、1 つの文書に含まれる固有表現 dNE の数に応じて、1 つの潜在表現空間 LS に存在するキーベクトル KV の数が、増加する。

【0126】

これらの結果として、本実施形態の情報処理装置 1 は、検索システム 200 における検索クエリに対する検索の精度を向上できる。

それゆえ、本実施形態の情報処理装置 1 は、タスクの信頼性を向上できる。

【0127】

(b) 情報処理方法

図 6 乃至図 14 を参照して、本実施形態の情報処理装置 1 による情報処理方法について、説明する。

尚、実施形態の情報処理方法は、実施形態の情報処理装置の制御方法を含み得る。

【0128】

< キーベクトル KV の生成 >

図 6 乃至図 8 を参照して、本実施形態の情報処理装置 1 における、検索システム 200

10

20

30

40

50

によるキーベクトル K V の生成処理について、説明する。

図 6 は、本実施形態の情報処理装置 1 におけるキーベクトル K V の生成処理を示すフローチャートである。図 7 及び図 8 のそれぞれは、本実施形態における、キーベクトル K V の生成処理を説明するための模式図である。

【0129】

< S 1 >

図 6 及び図 7 に示されるように、検索システム 200 において、抽出モジュール 210 は、プロセッサ 11 のインターフェース部 119 を介して、ストレージ 13 のデータベース 131 から複数の文書 D c を受ける。文書 D c は、1 つ以上の文を含む。複数の文書 D c は、情報処理装置 1 の外部のデバイスから提供されてもよい。

10

【0130】

例えば、抽出モジュール 210 は、文書 D c < 1 > として、“Biden was born in Scranton.” という文を、受ける。抽出モジュール 210 は、文書 D c < 2 > として、“Obama was born in Hawaii.” という文を、受ける。抽出モジュール 210 は、文書 D c < n > として、“Obama was awarded the Nobel peace prize.” という文を、受ける。

【0131】

< S 2 >

抽出モジュール 210 は、文書 D c と共に、文書 D c に関連付けられた識別番号 I d を受ける。抽出モジュール 210 は、識別番号 I d を抽出する。例えば、抽出モジュール 210 は、文書 D c から識別番号 I d を分離する。抽出された識別番号 I d は、テーブル管理部 290 に送られる。

20

【0132】

< S 3 >

抽出モジュール 210 は、固有表現抽出モデル M 1 に基づいて、各文書 D c に含まれる 1 つ以上の固有表現 d N E を抽出する。

図 7 の例において、抽出モジュール 210 は、文書 D c < 1 > から、“Biden” という固有表現 d N E a と、“Scranton” という固有表現 d N E b と、を抽出する。抽出モジュール 210 は、文書 D c < 2 > から、“Obama” という固有表現 d N E c と、“Hawaii” という固有表現 d N E d と、を抽出する。抽出モジュール 210 は、文書 D c < n > から、“Obama” という固有表現 d N E e と、“Nobel peace prize” という固有表現 d N E 30 f と、を抽出する。

【0133】

抽出モジュール 210 は、抽出された固有表現 d N E のそれぞれを、対応するエンコーダ 221 に送る。また、抽出モジュール 210 は、文書 D c から抽出された固有表現 d N E と共に、文書 D c 自体を、対応するエンコーダ 221 に送る。

【0134】

< S 4 >

エンコーダ 221 は、学習済み言語モデル M 2 に基づいて、キーベクトル K V を生成する。

【0135】

図 8 に示されるように、各エンコーダ 221 は、文書 D c 及び抽出された 1 つ以上の固有表現 d N E を受ける。

40

【0136】

各エンコーダ 221 は、受け取った文書 D c をトークン化する。各エンコーダ 221 は、トークン及び固有表現 d N E の埋め込み処理を行う。これによって、文書 D c を構成する語句及び固有表現 d N E は、ベクトル化される。

【0137】

エンコーダ 221 は、学習済み言語モデル M 2 に基づいて、或る文書 D c から抽出された固有表現 d N E 毎に、その文書 D c と固有表現 d N E とに関するキーベクトル K V を計算する。

50

【0138】

これによって、1つの文書 D_c に対して、その文書 D_c から抽出された固有表現 d_{NE} の数に応じて、1つ以上のキーベクトル KV が、生成される。

【0139】

例えば、文書 $D_c <1>$ に関して、2つのキーベクトル $KV <1> - a$, $KV <1> - b$ が、生成される。キーベクトル $KV <1> - a$ が、“Biden”という固有表現 $d_{NE a}$ と文書 $D_c <1>$ とによって、生成される。キーベクトル $KV <1> - b$ が、“Scranton”という固有表現 $d_{NE b}$ と文書 $D_c <1>$ とによって、生成される。

【0140】

例えば、文書 $D_c <2>$ に関して、2つのキーベクトル $KV <2> - c$, $KV <2> - d$ が、生成される。キーベクトル $KV <2> - c$ が、“Obama”という固有表現 $d_{NE c}$ と文書 $D_c <2>$ とによって、生成される。キーベクトル $KV <2> - d$ が、“Hawaii”という固有表現 $d_{NE d}$ と文書 $D_c <2>$ とによって、生成される。

10

【0141】

例えば、文書 $D_c <n>$ に関して、2つのキーベクトル $KV <n> - e$, $KV <n> - f$ が、生成される。キーベクトル $KV <n> - e$ が、“Obama”という固有表現 $d_{NE e}$ と文書 $D_c <n>$ とによって、生成される。キーベクトル $KV <n> - f$ が、“Nobel peace prize”という固有表現 $d_{NE f}$ と文書 $D_c <n>$ とによって、生成される。

【0142】

この結果として、知識源としてデータベース131に格納されている複数の文書 D_c に関する潜在表現空間 LS が生成される。複数の文書 D_c の複数のキーベクトル KV を含む潜在表現空間 LS は、同一の自然言語モデルを用いたニューラルネットワークに基づいて生成される。

20

【0143】

潜在表現空間 LS のキーベクトル KV の数は、データベース131に格納されている文書 D_c の数より多い。但し、各文書 D_c に含まれる固有表現 d_{NE} の数に応じて、キーベクトル KV の数が、文書 D_c の数より少なくなる場合もある。

【0144】

<S5>

検索システム200は、プロセッサ11によって、生成されたキーベクトル KV とキーベクトル KV に対応する識別番号 Id とを、ストレージ13に記憶する。これによって、潜在表現空間 LS は、ストレージ13に格納される。例えば、キーベクトル KV 及び識別番号 Id は、テーブル TBL において互いに関連付けられるように、保存される。

30

【0145】

例えば、テーブル管理部290は、キーベクトル KV 及び識別番号 Id を受ける。テーブル管理部290は、キーベクトル KV 及び識別番号 Id の対応関係を示すテーブル TBL を生成する。テーブル TBL は、検索タスクの実行前に生成されていればよい。

【0146】

例えば、検索タスクの実行時、生成された複数のキーベクトル KV は、メモリ12に読み出される。検索タスクの実行時、キーベクトル KV 及び識別番号 Id に関するテーブル TBL は、テーブル管理部290に読み出される。尚、キーベクトル KV は、テーブル管理部290に読み出されてもよい。

40

【0147】

以上のように、複数の文書 D_c から、キーベクトル KV が生成される。

これによって、複数のキーベクトル KV を含む潜在表現空間 LS が、形成される。

【0148】

<検索タスク>

図9乃至図13を参照して、本実施形態の情報処理装置1における、検索システム200による検索タスクについて、説明する。

図9は、本実施形態の情報処理装置1における検索タスクを示すフローチャートである

50

。図 10 乃至図 13 のそれぞれは、本実施形態における、検索タスクを説明するための模式図である。

【0149】

<S10>

図 9 に示されるように、検索システム 200 は、検索タスクの対象となる質問 Qs を受ける。質問 Qs は、情報処理装置 1 のユーザー又はデバイス 9 によって生成される。ユーザーは、質問 Qs を、ユーザーインターフェース 14 を介して、情報処理装置 1 に供給する。質問 Qs は、プロセッサ 11 のインターフェース部 119 を介して、検索システム 200 に入力される。質問 Qs は、1 つ以上の文を含む。

【0150】

図 10 に示されるように、検索システム 200 において、抽出モジュール 210 は、質問 Qs を受ける。

【0151】

例えば、本実施形態において、抽出モジュール 210 は、“Where was Obama born?” という文を、質問 Qs として、受ける。

【0152】

<S11>

抽出モジュール 210 は、固有表現抽出モデル M1 に基づいて、質問 Qs から固有表現 qNE を抽出する。

図 10 の例において、抽出モジュール 210 は、質問 Qs から “Obama” という固有表現 qNE を抽出する。

【0153】

抽出モジュール 210 は、抽出された固有表現 qNE を、エンコーダ 220 に送る。抽出モジュール 210 は、質問 Qs から抽出された固有表現 qNE と共に、質問 Qs 自体を、エンコーダ 220 に送る。

【0154】

<S12>

エンコーダ 220 は、質問 Qs 及び固有表現 qNE を受ける。

エンコーダ 220 は、学習済み言語モデル M2 に基づいて、質問 Qs 及び固有表現 qNE を用いて、クエリベクトル QV を計算する。

これによって、質問 Qs に対応するクエリベクトル QV が、生成される。生成されたクエリベクトル QV は、質問 Qs の特徴量ベクトルに質問 Qs から抽出された固有表現 qNE の特徴量ベクトルが反映された値を有する。

【0155】

<S13>

図 11 に示されるように、検索システム 200 は、質問 Qs の受信及び質問 Qs に対する各種の処理に並行して、プロセッサ 11 の制御によって、潜在表現空間 LS に含まれる複数のキーベクトル KV を、ストレージ 13 のデータベース 132 から読み出す。読み出されたキーベクトル KV は、例えば、メモリ 12 に格納される。

【0156】

検索システム 200 は、プロセッサ 11 によって、識別番号 Id 及びキーベクトル KV に関するテーブル TBL を、テーブル管理部 290 に読み出す。

【0157】

尚、キーベクトル KV 及びテーブル TBL は、質問 Qs の受信の前に、ストレージ 13 から読み出されていてもよい。

【0158】

<S14>

図 12 に示されるように、類似度計算部 230 は、クエリベクトル QV と各キーベクトル KV との間の類似度を計算する。例えば、クエリベクトル QV とキーベクトル KV との間の類似度は、コサイン類似度によって、計算される。

10

20

30

40

50

【0159】

これによって、クエリベクトルQVと各キーベクトルKVとの間の類似度の計算結果を示すスコアSCRが、キーベクトルKVごとにそれぞれ得られる。

【0160】

類似度計算部230は、スコアSCRを判定部240に送る。

【0161】

<S15>

判定部240は、類似度の計算結果に基づいて、回答候補となり得るキーベクトルKVの判定処理を、実行する。

【0162】

図13に示されるように、判定部240は、類似度のスコアSCRを、類似度計算部230から受ける。判定部240は、クエリベクトルQVに類似するキーベクトルKVを、潜在表現空間LSの複数のキーベクトルKVの中から検索する。

判定部240は、各キーベクトルKVに関するスコアSCRに基づいて、質問Qsに対する回答候補となり得るキーベクトルKVを、複数のキーベクトルKVの中から判定する。

【0163】

例えば、判定部240は、類似度のスコアSCRが閾値を満たすか否かに応じて、クエリベクトルQVに対するキーベクトルKVのランク付けを行う。これによって、判定部240は、或る閾値以上の類似度を有するキーベクトルKVを、回答候補として取得する。

【0164】

図13の例において、判定部240は、文書Dc自体の文脈の特徴と共に“Obama”という固有表現dNEの特徴が反映された複数のキーベクトルKV<2>-c, KV<n>-eを選択し、取得する。

【0165】

<S16>

図13に示されるように、結果出力部250は、判定処理の結果に基づいて、1つ以上の文書Dcを、質問Qsに対する検索結果RRとして、取得及び出力する。

【0166】

例えば、結果出力部250は、類似度に基づくキーベクトルKVの判定結果を判定部240から受ける。

【0167】

結果出力部250は、受けたキーベクトルKVに基づいて、テーブル管理部290のテーブルTBLを参照する。結果出力部250は、テーブルTBLの参照結果に基づいて、回答候補のキーベクトルKVに対応する識別番号Idを検知する。

【0168】

結果出力部250は、ストレージ13のデータベース131にアクセスする。結果出力部250は、検知された識別番号Idに対応する文書Dcを、データベース131から読み出す。

【0169】

例えば、キーベクトルKV<2>-cが選択された場合、結果出力部250は、選択されたキーベクトルKV<2>-c及びテーブルTBLに基づいて、キーベクトルKV<2>-cに識別番号Id<2>が関連付けられていることを、検知する。

これによって、結果出力部250は、検知された識別番号Id<2>に対応する文書Dc<2>を、データベース131から読み出す。

【0170】

キーベクトルKV<n>-eが選択された場合、結果出力部250は、選択されたキーベクトルKV<n>-e及びテーブルTBLに基づいて、キーベクトルKV<n>-eに識別番号Id<n>が関連付けられていることを、検知する。

これによって、結果出力部250は、検知された識別番号Id<n>に対応する文書D

10

20

30

40

50

$c < n >$ を、データベース131から読み出す。

【0171】

このように、結果出力部250は、回答候補として選択されたキーベクトルKVに対応する文書Dc ($Dc < 2 >$, $Dc < n >$)を、取得する。

【0172】

結果出力部250は、取得された文書Dc $< 2 >$, $Dc < n >$ を、回答に関するデータアイテムとして、ユーザーインターフェース14に出力する。

【0173】

この結果として、質問Qsを提供したユーザーは、結果出力部250からの出力に基づいて、質問Qsに対する検索結果RRの文書Dcを、認識できる。

【0174】

以上の処理によって、本実施形態の情報処理装置1における、検索タスクが、完了する。

【0175】

(c) まとめ

DPR (Dense passage retrieval) のような深層学習を用いた文書検索方法は、文字列の一致による関連文書の検索方法のような古典的な検索方法に比較して、DPRとして機能するために必要な追加学習に用いたデータセットに対してより高精度の検索を実現できる場合がある。

【0176】

但し、DPRによる文書検索方法は、初見の文書に対する検索精度が低い傾向がある。例えば、固有表現に関する質問に対する検索タスクにおいて、DPRによる文書検索方法の検索性能は、古典的な検索方法の検索性能より低いことがある。

【0177】

上述のように、本実施形態の情報処理装置1において、検索システム200は、質問Qs及び検索対象の文書Dcのそれぞれから、固有表現qNE, dNEを抽出する。

【0178】

検索システム200は、質問Qsと抽出された固有表現qNEとからクエリベクトルQVを生成する。検索システム200は、文書Dcから抽出された複数の固有表現dNEごとに、文書Dcと各固有表現dNEとから複数のキーベクトルKVを生成する。

【0179】

これによって、本実施形態の情報処理装置1は、同一の潜在表現空間LS上に、文書Dcの数より多い数のキーベクトルKVを、存在させることができる。また、本実施形態において、1つの文書Dcに対するキーベクトルKVの冗長性が、高くなる。

【0180】

本実施形態において、質問Qsに対応するクエリベクトルQVは、質問Qsの文脈の特徴に加えて、質問Qsから抽出された固有表現qNEの特徴がさらに反映される。これと同様に、文書Dcに対応するキーベクトルKVは、文書Dcの文脈の特徴に加えて、文書Dcから抽出された固有表現dNEの特徴がさらに反映される。

これによって、本実施形態の情報処理装置1は、質問Qsに含まれる固有表現qNEに類似する固有表現dNEの特徴が反映されたキーベクトルKVを、複数のキーベクトルKVの中から検索できる。

これらの結果として、本実施形態の情報処理装置1において、初見の文に対する検索性能が向上する。

【0181】

図14は、本実施形態の情報処理装置1の検索タスクに関する特性を示すグラフである。

図14のグラフの横軸は、検索結果として取得され得るキーベクトルの個数に対応する。図14のグラフの縦軸は、検索精度 (単位: %) に対応する。

【0182】

丸のプロット及び実線で示される特性 A 1 は、実施形態の情報処理装置 1 の検索システム 200 における検索精度の推移を示している。三角のプロット及び破線で示される特性 A 2 は、比較例の検索システムにおける検索精度の推移を示している。比較例の検索システムは、一般的な DPR を用いた検索システムである。

【0183】

本実施形態の検索システム 200 の特性及び比較例の検索システムの特性は、Entity Questions データセットを用いて評価された。開発セットにおける質問の数は、10170 問である。

【0184】

図 14 に示されるように、本実施形態のように抽出された固有表現 dNE と文書 Dc とから生成されたキーベクトル KV が、検索システム 200 の検索タスクに用いられた場合、本実施形態の検索システム 200 の検索精度は、比較例の検索システムの検索精度に比較して、高くなる。

【0185】

このように、本実施形態の情報処理装置 1 は、検索タスクの検索精度を向上できる。

【0186】

以上のように、本実施形態の情報処理装置 1 及び情報処理方法は、タスクの精度を向上できる。

【0187】

(2) 変形例

図 15 を参照して、実施形態の情報処理装置 1 の変形例について、説明する。

【0188】

図 15 は、実施形態の情報処理装置 1 の変形例を説明するための模式図である。

【0189】

図 15 に示されるように、質問 Qsx が、複数の固有表現 qNEa, qNEb を含む場合がある。例えば、図 15 の例において、“Which Nobel prize was awarded to Obama?” という質問 Qsx が、情報処理装置 1 の検索システム 200 に供給される。質問 Qs は、“Obama” という固有表現 qNEa と、“Nobel prize” という固有表現 qNEb と、を含む。

【0190】

この場合において、文書 Dc 及びキーベクトル KV と同様に、複数のクエリベクトル QVa, QVb が、1 つの質問 Qsx に含まれる複数の固有表現 qNE 毎に、計算及び生成される。

【0191】

抽出モジュール 210 は、固有表現抽出モデル M1 に基づいて、複数の固有表現 qNEa, qNEb を、質問 Qsx から抽出する。ここでは、“Obama” と “Nobel prize” という 2 つの固有表現 qNEa, qNEb が、それぞれ抽出される。

【0192】

抽出モジュール 210 は、質問 Qsx、及び、抽出された 2 つの固有表現 qNEa, qNEb を、エンコーダ 220 へ送る。

【0193】

エンコーダ 220 は、学習済み言語モデル M2 に基づいて、固有表現 qNEa, qNEb 毎に、質問 Qsx と各固有表現 qNEa, qNEb とを用いたクエリベクトル QVa, QVb を計算する。これによって、2 つのクエリベクトル QVa, QVb が、1 つの質問 Qsx から生成される。

【0194】

一方のクエリベクトル QVa に、質問 Qsx の文脈の特徴に対して固有表現 qNEa の特徴が反映されている。他方のクエリベクトル QVb に、質問 Qsx の文脈の特徴に対して固有表現 qNEb の特徴が反映されている。

【0195】

10

20

30

40

50

類似度計算部 230 は、複数のクエリベクトル $Q V a$, $Q V b$ と複数のキーベクトル $K V$ との間の類似度を、それぞれ計算する。

【0196】

類似度計算部 230 は、クエリベクトル $Q V a$ と各キーベクトル $K V$ との間の類似度を計算する。これによって、クエリベクトル $Q V a$ 及び各キーベクトル $K V$ に関する類似度の計算結果として、複数のスコア $S C R a$ が、得られる。

類似度計算部 230 は、クエリベクトル $Q V b$ と各キーベクトル $K V$ との間の類似度を計算する。これによって、クエリベクトル $Q V b$ 及び各キーベクトル $K V$ に関する類似度の計算結果として、複数のスコア $S C R b$ が、得られる。

【0197】

判定部 240 は、各クエリベクトル $Q V a$, $Q V b$ に関する類似度の計算結果に基づいて、潜在表現空間 $L S$ の複数のキーベクトル $K V$ の中から質問 $Q s x$ に対する回答候補となり得るキーベクトル $K V$ を、判定する。

【0198】

結果出力部 250 は、判定結果及びテーブル $T B L$ に基づいて、回答となる文書 $D c$ を、データベース 131 から検索する。これによって、結果出力部 250 は、1 以上の文書 $D c$ をデータベース 131 から取得する。

結果出力部 250 は、検索結果 $R R$ の文書 $D c$ を出力する。

【0199】

本変形例のように、ユーザーからの質問 $Q s x$ が、複数の固有表現 $q N E a$, $q N E b$ を含む場合であっても、実施形態の検索システム 200 は、固有表現 $q N E a$, $q N E b$ 毎に生成された複数のクエリベクトル $Q V a$, $Q V b$ を用いて、質問 $Q s x$ に対する回答となる文書 $D c$ を、取得できる。

【0200】

したがって、本変形例の情報処理装置 1 は、上述の実施形態と実質的に同じ効果を得ることができる。

【0201】

(3) その他

本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。

【符号の説明】

【0202】

1：情報処理装置、11：プロセッサ、12：メモリ、13：ストレージ、14：ユーザーインターフェース、200：検索システム、210：抽出モジュール、220、221：エンコーダ、230：類似度計算部、240：判定部、250：結果出力部。

10

20

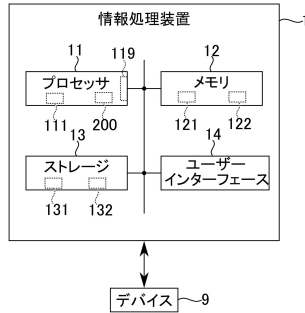
30

40

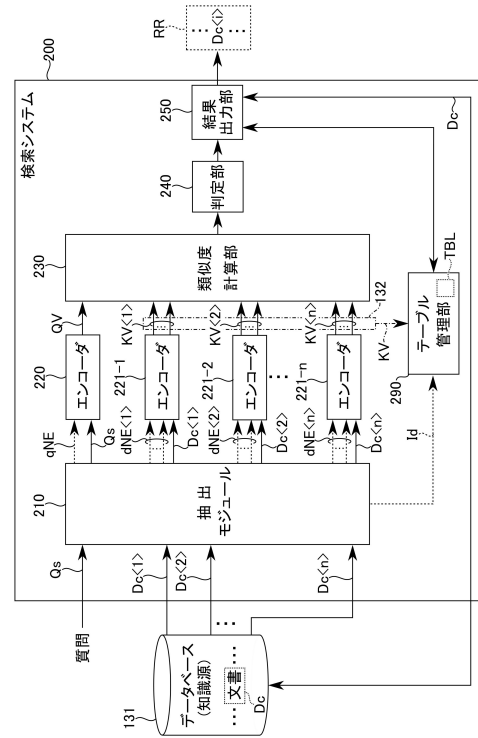
50

【図面】

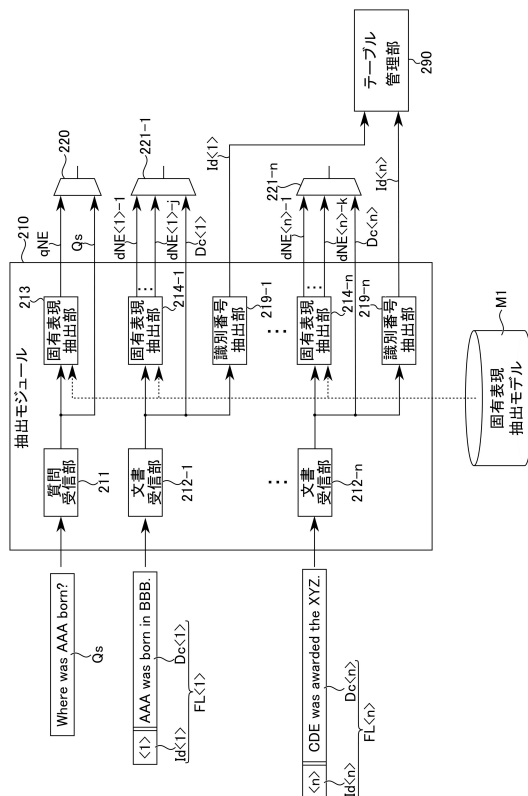
【図 1】



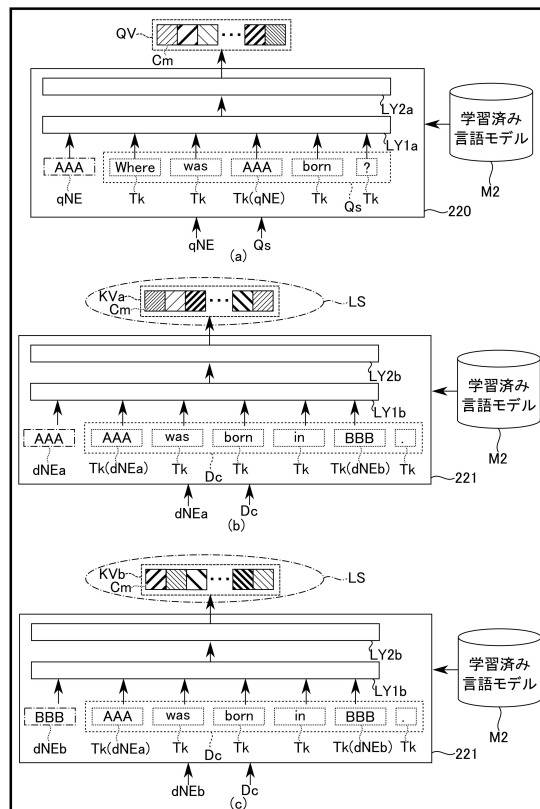
【図 2】



【図 3】



【図 4】



10

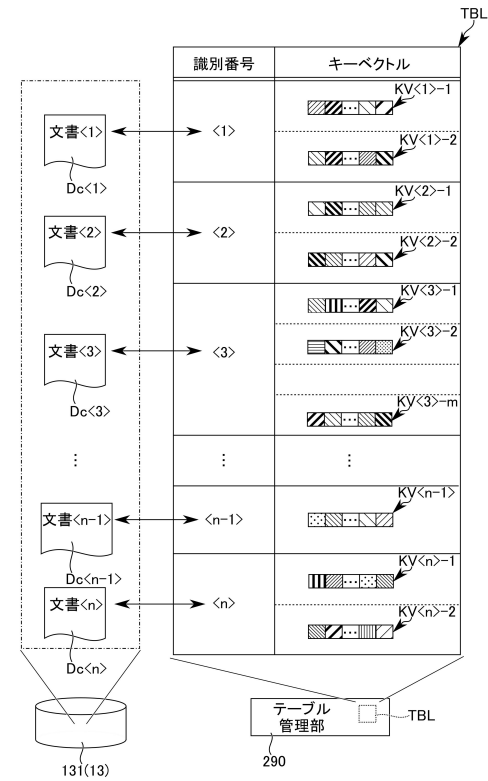
20

30

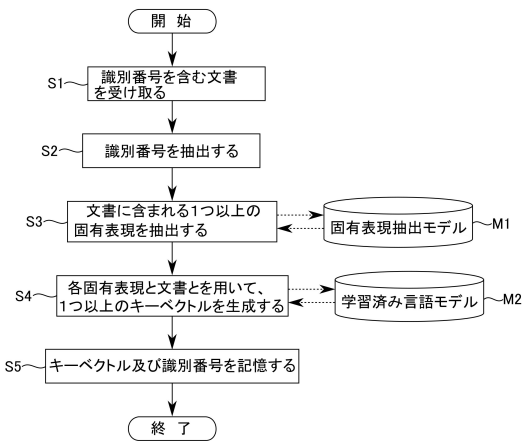
40

50

【図 5】



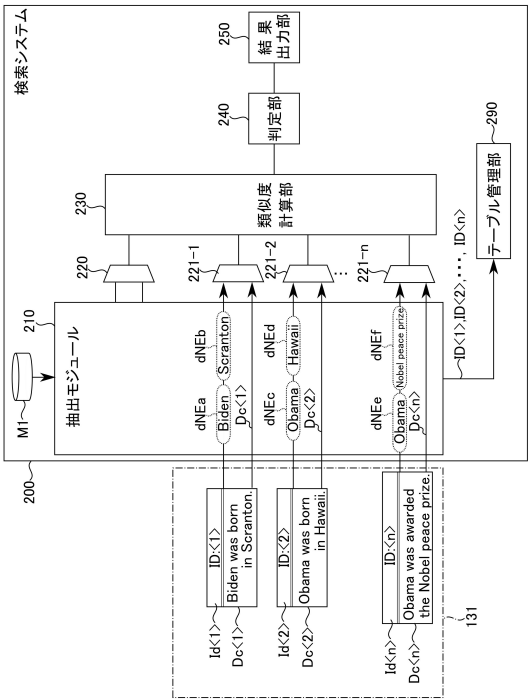
【図 6】



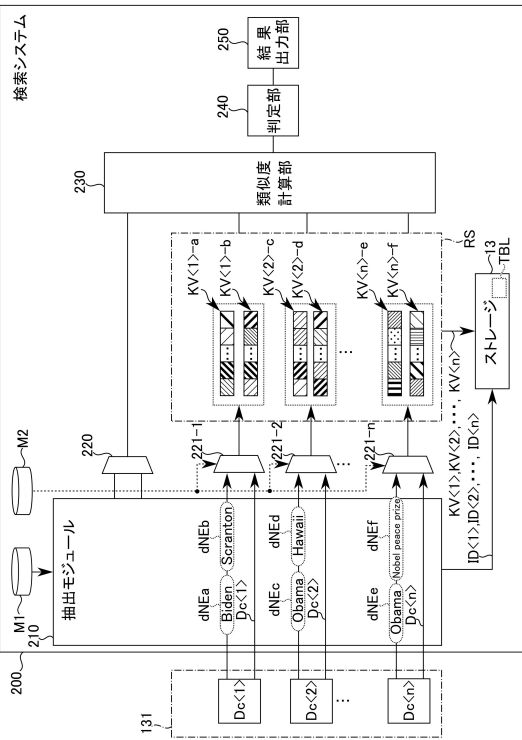
10

20

【図 7】



【図 8】

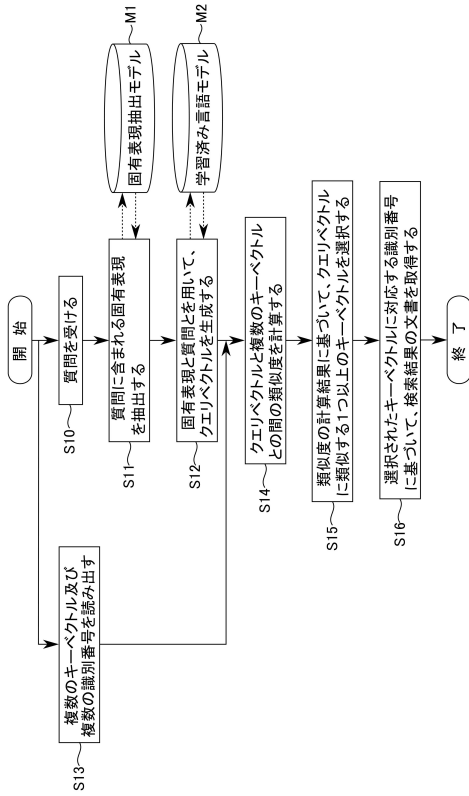


30

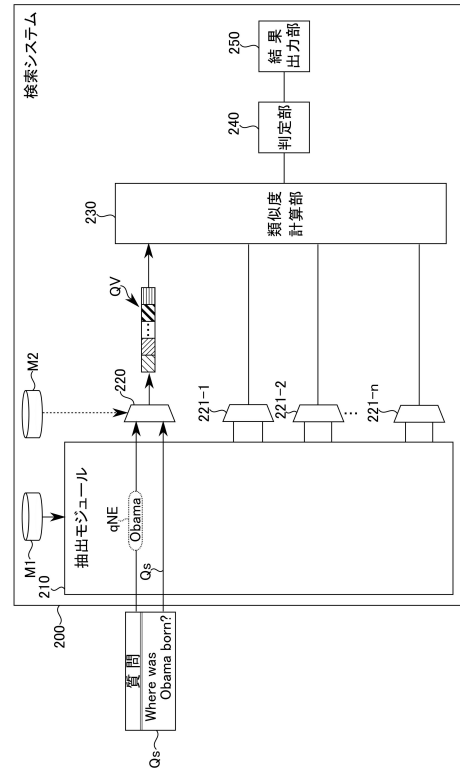
40

50

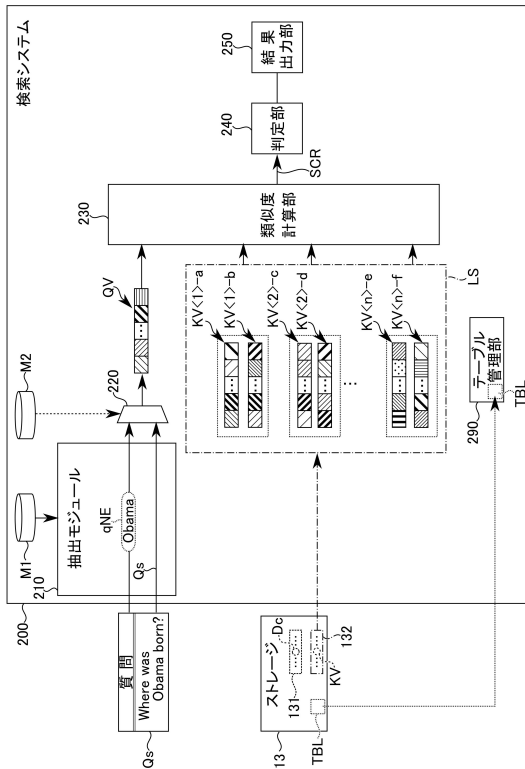
【圖 9】



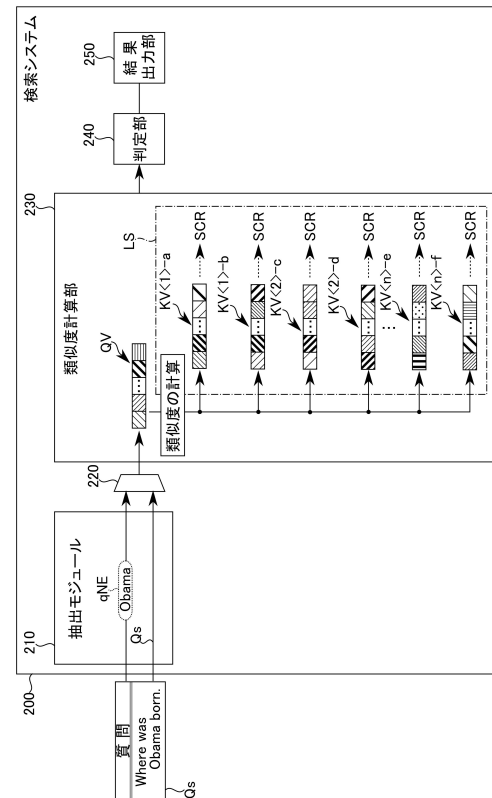
【圖 10】



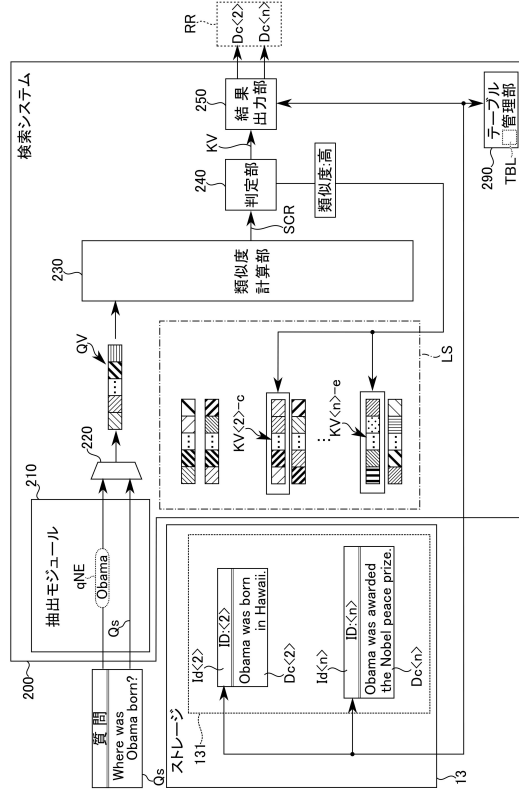
【図 1 1】



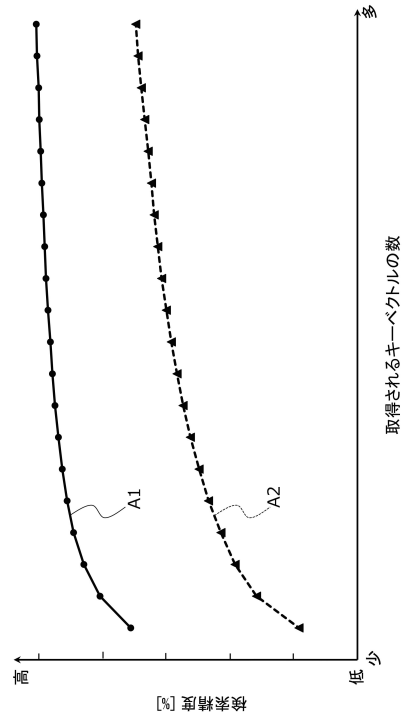
【圖 12】



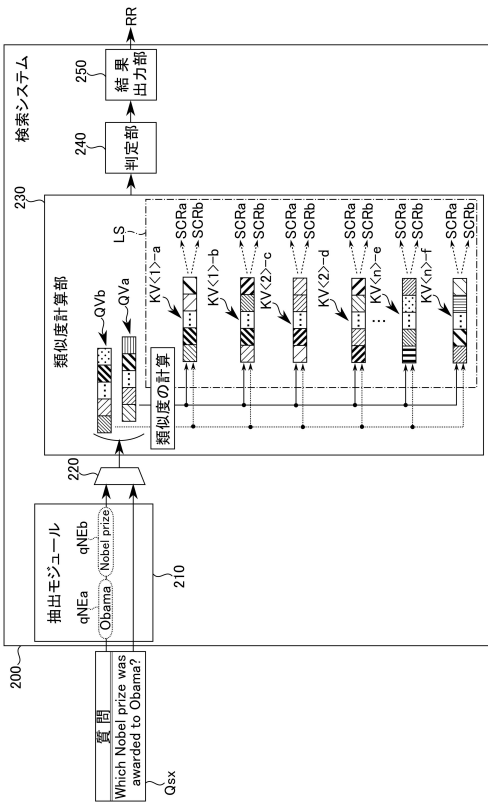
【図 13】



【図 14】



【図 15】



10

20

30

40

50