(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7547653号
(P7547653)

(45)発行日　令和6年9月9日(2024.9.9)

(24)登録日　令和6年8月30日(2024.8.30)

(51)国際特許分類

Ｇ０６Ｆ　40/56　(2020.01)		Ｇ０６Ｆ	40/56
Ｇ０６Ｆ　40/44　(2020.01)		Ｇ０６Ｆ	40/44

ＦＩ

請求項の数　19　（全24頁）

(21)出願番号　特願2023-561822(P2023-561822)
(86)(22)出願日　令和4年5月20日(2022.5.20)
(65)公表番号　特表2024-520994(P2024-520994A)
(43)公表日　令和6年5月28日(2024.5.28)
(86)国際出願番号　PCT/US2022/030249
(87)国際公開番号　WO2022/246194
(87)国際公開日　令和4年11月24日(2022.11.24)
審査請求日　令和5年11月21日(2023.11.21)
(31)優先権主張番号　63/191,563
(32)優先日　令和3年5月21日(2021.5.21)
(33)優先権主張国・地域又は機関
米国(US)
早期審査対象出願

(73)特許権者　502208397
グーグル エルエルシー
Ｇｏｏｇｌｅ ＬＬＣ
アメリカ合衆国 カリフォルニア州 ９４０４３ マウンテン ビュー アンフィシアター パークウェイ １６００
１６００ Ａｍｐｈｉｔｈｅａｔｒｅ Ｐａｒｋｗａｙ ９４０４３ Ｍｏｕｎｔａｉｎ Ｖｉｅｗ， ＣＡ Ｕ．Ｓ．Ａ．
(74)代理人　100108453
弁理士　村山 靖彦
(74)代理人　100110364
弁理士　実広 信哉
(74)代理人　100133400
弁理士　阿部 達彦

最終頁に続く

(54)【発明の名称】　文脈テキスト生成のサービスにおいて中間テキスト分析を生成する機械学習済み言語モデル

(57)【特許請求の範囲】

【請求項１】

改善された解釈可能性を有する文脈テキスト生成のためのコンピューティングシステムであって、

1つまたは複数のプロセッサと、

1つまたは複数の非一時的コンピュータ可読媒体とを含み、前記1つまたは複数の非一時的コンピュータ可読媒体が、

文脈テキスト生成のサービスにおいてテキスト分析を実行する機械学習済み言語モデルと、

前記1つまたは複数のプロセッサによって実行されるときに前記コンピューティングシステムに動作を実行させる命令とをまとめて記憶し、前記動作が、

1つまたは複数の文脈テキストトークンを含む文脈テキスト文字列を取得することと、

1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記文脈テキスト文字列を処理することであって、前記1つまたは複数の中間テキストトークンが、前記文脈テキスト文字列に含まれない追加の情報にアクセスするために構造ツールの使用を引き起こす少なくとも1つのツールトークンを含む、ことと、

1つまたは複数の出力テキストトークンを含む出力テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記1つまたは複数の中間テキスト文字列を処

理することとを含み、
前記1つまたは複数の中間テキスト文字列は、前記出力テキスト文字列をサポートする前記文脈テキスト文字列のテキスト分析を含む、コンピューティングシステム。
【請求項２】
前記構造ツールが、データベースから追加の情報にアクセスするためにデータベース検索を含む、請求項1に記載のコンピューティングシステム。
【請求項３】
前記構造ツールが、アプリケーションプログラミングインターフェース(API)を介して追加の情報を要求して受信するために、API呼び出しを含む、請求項1に記載のコンピューティングシステム。
【請求項４】
前記構造ツールが、入力テキストトークン上で1つまたは複数の動作のシーケンスを実行するプログラミング言語インタープリタを含む、請求項1に記載のコンピューティングシステム。
【請求項５】
前記構造ツールが、検索エンジン、ナレッジグラフ、またはデジタルアシスタントからの結果をクエリするクエリサービスを含む、請求項1に記載のコンピューティングシステム。
【請求項６】
前記機械学習済み言語モデルが前記ツールトークンを生成するとき、前記動作が、
前記機械学習済み言語モデルを中断することと、
前記追加の情報にアクセスするために前記構造ツールを実行することと、
前記1つまたは複数の中間テキスト文字列の現在のバージョンに前記追加の情報を付加することと、
前記1つまたは複数の中間テキスト文字列の前記現在のバージョンおよび前記付加された追加の情報に基づいて前記機械学習済み言語モデルを用いてテキスト生成を再開することとを含む、請求項1に記載のコンピューティングシステム。
【請求項７】
オーディオ出力を生成するために音声合成システムを用いて出力テキスト文字列を処理することをさらに含む、請求項1に記載のコンピューティングシステム。
【請求項８】
前記機械学習済み言語モデルは、トークンごとに動作し、前記1つまたは複数の中間テキスト文字列を生成するとき、各生成された中間テキストトークンを再帰的に入力として受信する、請求項1に記載のコンピューティングシステム。
【請求項９】
1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列を生成するために前記機械学習済み言語モデルを用いて前記文脈テキスト文字列を処理することが、
第1の反復に対して、
1つまたは複数の中間テキストトークンを含む第1の中間テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記文脈テキスト文字列を処理することと、
更新された文脈テキスト文字列を生成するために、前記文脈テキスト文字列に前記第1の中間テキスト文字列を付加することと、
1つまたは複数の追加の反復の各々に対して、前記機械学習済み言語モデルが閉止トークンを出力するまで、
1つまたは複数の中間テキストトークンを含む追加の中間テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記更新された文脈テキスト文字列を処理することと、
次の反復に対して前記更新された文脈テキスト文字列を生成するために、前記更新された文脈テキスト文字列に前記追加の中間テキスト文字列を付加することとを含む、請求

項1に記載のコンピューティングシステム。
【請求項１０】
前記機械学習済み言語モデルが、複数の訓練タプル上で訓練されており、各訓練タプルは、例示的な文脈テキスト文字列、1つまたは複数の例示的な中間テキスト文字列、および例示的な出力テキスト文字列を含む、請求項1に記載のコンピューティングシステム。
【請求項１１】
少なくとも前記1つまたは複数の例示的な中間テキスト文字列が、ヒューマンラベラーによって生成されている、請求項10に記載のコンピューティングシステム。
【請求項１２】
前記機械学習済み言語モデルが質問応答モデルを含み、前記文脈テキスト文字列が質問を含む、請求項1に記載のコンピューティングシステム。
【請求項１３】
前記機械学習済み言語モデルが対話モデルを含み、前記文脈テキスト文字列が対話履歴を含む、請求項1に記載のコンピューティングシステム。
【請求項１４】
前記機械学習済み言語モデルが、
回帰型ニューラルネットワーク、
マルチヘッドセルフアテンションモデル、または
シーケンスツーシーケンスモデルを含む、請求項1に記載のコンピューティングシステム。
【請求項１５】
前記文脈テキスト文字列の少なくとも一部が、ユーザによって入力されたテキストを含み、前記動作が、前記ユーザに表示するために少なくとも前記出力テキスト文字列を提供することをさらに含む、請求項1に記載のコンピューティングシステム。
【請求項１６】
前記文脈テキスト文字列が、中間テキスト文字列を生成することなく、元の文脈テキスト文字列からベース出力を直接生成するように構成された機械学習済み言語モデルによって生成された前記ベース出力と連結された前記元の文脈テキスト文字列を含む、請求項1に記載のコンピューティングシステム。
【請求項１７】
改善された文脈テキスト生成のためのコンピュータ実装方法であって、
複数の訓練タプルを取得するステップであって、各訓練タプルは、1つまたは複数の文脈テキストトークンを含む例示的な文脈テキスト文字列、1つまたは複数の中間テキストトークンを含む1つまたは複数の例示的な中間テキスト文字列、および1つまたは複数の出力テキストトークンを含む例示的な出力テキスト文字列を含む、ステップと、
各訓練タプルに対して、
前記訓練タプルの少なくとも一部を言語モデルに入力するステップと、
予測された次のトークンを前記言語モデルの出力として受信するステップであって、前記予測される次のトークンは、前記訓練タプルの前記一部を処理することによって前記言語モデルによって生成される、ステップと、
前記言語モデルによって生成された前記予測される次のトークンを、前記訓練タプルに含まれる実際の次のトークンと比較する損失関数を評価するステップと、
前記損失関数の前記評価に基づいて前記言語モデルの1つまたは複数のパラメータの1つまたは複数の値を修正するステップとを含み、
前記訓練タプルのうちの少なくとも1つに対して、前記1つまたは複数の中間テキストトークンが、前記例示的な文脈テキスト文字列に含まれない追加の情報にアクセスするために、構造ツールの使用を引き起こす少なくとも1つのツールトークンを含み、前記追加の情報が、前記1つまたは複数の中間テキストトークンに含まれる、コンピュータ実装方法。
【請求項１８】
各訓練タプルに対して、前記入力するステップ、受信するステップ、評価するステップ

、および修正するステップが、前記1つまたは複数の例示的な中間テキスト文字列および前記例示的な出力テキスト文字列に含まれる各トークンに対して実行される、請求項17に記載のコンピュータ実装方法。

【請求項１９】

改善された解釈可能性を有する文脈テキスト生成のためのコンピューティングシステムであって、

1つまたは複数のプロセッサと、

1つまたは複数の非一時的コンピュータ可読媒体とを含み、前記1つまたは複数の非一時的コンピュータ可読媒体が、

文脈テキスト生成のサービスにおいてテキスト分析を実行する機械学習済み言語モデルと、

前記1つまたは複数のプロセッサによって実行されるときに前記コンピューティングシステムに動作を実行させる命令とをまとめて記憶し、前記動作が、

1つまたは複数の文脈テキストトークンを含む文脈テキスト文字列を取得することと、

1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記文脈テキスト文字列を処理することであって、1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列を生成するために前記機械学習済み言語モデルを用いて前記文脈テキスト文字列を処理することが、

第1の反復に対して、

1つまたは複数の中間テキストトークンを含む第1の中間テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記文脈テキスト文字列を処理することと、

更新された文脈テキスト文字列を生成するために、前記文脈テキスト文字列に前記第1の中間テキスト文字列を付加することと、

1つまたは複数の追加の反復の各々に対して、前記機械学習済み言語モデルが閉止トークンを出力するまで、

1つまたは複数の中間テキストトークンを含む追加の中間テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記更新された文脈テキスト文字列を処理することと、

次の反復に対して前記更新された文脈テキスト文字列を生成するために、前記更新された文脈テキスト文字列に前記追加の中間テキスト文字列を付加することとを含む、ことと、

1つまたは複数の出力テキストトークンを含む出力テキスト文字列を生成するために、前記機械学習済み言語モデルを用いて前記1つまたは複数の中間テキスト文字列を処理することとを含み、

前記1つまたは複数の中間テキスト文字列は、前記出力テキスト文字列をサポートする前記文脈テキスト文字列のテキスト分析を含む、コンピューティングシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

関連出願

本出願は、2021年5月21日に出願された米国仮特許出願第63/191,563号の優先権および利益を主張する。米国仮特許出願第63/191,563号は、その全体が参照により本明細書に組み込まれる。

【０００２】

本開示は、一般に、言語モデリングに対する機械学習の使用に関する。より詳細には、本開示は、文脈テキスト生成のサービスにおいて中間テキスト分析(たとえば、APIなどの構造ツールの使用を含む)を生成する機械学習済み言語モデルに関する。

【背景技術】

【０００３】

自然言語処理(NLP)は、近年急速な発展が見られ、そのような進歩は主に、学習ベースのアルゴリズムおよび機械学習または「ニューラル」学習の他の態様における改善に起因

する。NLPの分野における1つの特定のタスクは、文脈テキスト生成である。文脈テキスト生成タスクでは、エージェント(たとえば、機械学習モデル)が、与えられた文脈から出力テキストを生成する任務を負う。たとえば、与えられた文脈は、1つまたは複数の文脈テキスト文字列を含み得る。そのため、文脈テキスト生成タスクに対するいくつかの例示的な手法において、テキストツーテキストモデルが、入力文脈テキストを読み出し、次いで出力テキストを直接作成する。

【０００４】

文脈テキスト生成タスクの1つの例は、質問応答タスクである。質問応答タスクでは、質問が入力文脈であり、所望の出力が質問に対する応答である。文脈テキスト生成タスクの別の例は、対話生成である。対話生成では、入力文脈は会話履歴であり、所望の出力は次の発話であり、ここで次の発話は、会話履歴の文脈に応答するか、または場合によっては会話履歴の文脈の中で合理的である。

【０００５】

文脈テキスト生成に対する現在の最先端モデルは、GPT3(Brownら、Language Models are Few-Shot Learners, arXiv:2005.14165)のような左から右への言語モデル、ここで"〈input〉 〈output〉"は1つのシーケンスと見なされる、かまたは、元のトランスフォーマ(Vaswaniら、Attention is All You Need, arXiv:1706.03762)のようなシーケンスツーシーケンスモデルのいずれかの、トランスフォーマベースのニューラルモデルである傾向がある。

【０００６】

しかしながら、GPT3およびトランスフォーマなどのニューラル言語モデルは、いくつかの欠点の影響を受ける。具体的には、ニューラル言語モデルは著しい知性を示すが、それらの知識は、それらの訓練データセットの中に含まれる(およびそれらから学習される)情報および/または文脈テキスト入力の中で導入された情報に制約される。したがって、事実情報のそれらの知識は極めて限定され、一般に時間的に凍結されている。そのため、事実情報を含む出力を作成することを要求されるとき、モデルは、一般的に、間違った事実を錯覚するか、または古い情報を供給するかのいずれかである。間違った事実情報への依存は、非効率性をもたらすことがあり、そこにおいて、間違った行動(たとえば、コンピューター化された行動)が取られ、訂正または場合によっては修正される必要があり、冗長で不必要なリソース(たとえば、計算リソース)の使用をもたらす。

【０００７】

ニューラル言語モデルの別の例示的な欠点として、それらの出力は、解釈または理解することが困難である。具体的には、そのようなモデルはしばしば、入力から出力を直接生成するので、そのようなモデルがなぜその出力を生成したのか、または入力のどの態様がその出力につながったのかを正確に理解することは困難である。言語モデル出力における解釈可能性の欠如は、モデル出力に対する信頼または信用の欠如をもたらすことがあり、それは、モデルの出力の正確さまたは有用性を「ダブルチェック」しようとする、不必要なオーバーヘッドまたは他の努力(たとえば、コンピュータ化された動作)をもたらすことがある。

【先行技術文献】

【非特許文献】

【０００８】

【文献】Brownら、Language Models are Few-Shot Learners, arXiv:2005.14165
【文献】Vaswaniら、Attention is All You Need, arXiv:1706.03762

【発明の概要】

【課題を解決するための手段】

【０００９】

本開示の実施形態の態様および利点は、以下の説明において部分的に記載されるか、または説明から知ることができるか、または実施形態の実施を通して知ることができる。

【００１０】

本開示の1つの例示的な態様は、改善された解釈可能性を有する文脈テキスト生成のためのコンピューティングシステムを対象とする。コンピューティングシステムは、1つまたは複数のプロセッサと1つまたは複数の非一時的コンピュータ可読媒体とを含み、1つまたは複数の非一時的コンピュータ可読媒体は、文脈テキスト生成のサービスにおいてテキスト分析を実行する機械学習済み言語モデルと、1つまたは複数のプロセッサによって実行されるときにコンピューティングシステムに動作を実行させる命令とをまとめて記憶する。動作は、1つまたは複数の文脈テキストトークンを含む文脈テキスト文字列を取得することを含む。動作は、1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列を生成するために、機械学習済み言語モデルを用いて文脈テキスト文字列を処理することを含む。動作は、1つまたは複数の出力テキストトークンを含む出力テキスト文字列を生成するために、機械学習済み言語モデルを用いて1つまたは複数の中間テキスト文字列を処理することを含む。1つまたは複数の中間テキスト文字列は、出力テキスト文字列をサポートする文脈テキスト文字列のテキスト分析を含む。

【００１１】

本開示の別の例示的な態様は、改善された文脈テキスト生成のためのコンピュータ実装方法を対象とする。

【００１２】

方法は、複数の訓練タプルを取得するステップを含み、各訓練タプルは、1つまたは複数の文脈テキストトークンを含む例示的な文脈テキスト文字列と、1つまたは複数の中間テキストトークンを含む1つまたは複数の例示的な中間テキスト文字列と、1つまたは複数の出力テキストトークンを含む例示的な出力テキスト文字列とを含む。方法は、各訓練タプルに対して、訓練タプルの少なくとも一部を言語モデルに入力するステップと、予測される次のトークンを言語モデルの出力として受信するステップであって、予測される次のトークンは訓練タプルの一部を処理することによって言語モデルによって生成される、ステップと、言語モデルによって生成された予測される次のトークンを訓練タプルに含まれる実際の次のトークンと比較する損失関数を評価するステップと、損失関数の評価に基づいて言語モデルの1つまたは複数のパラメータの1つまたは複数の値を修正するステップとを含む。

【００１３】

本開示の他の態様は、様々なシステム、装置、非一時的コンピュータ可読媒体、ユーザインターフェース、および電子デバイスを対象とする。

【００１４】

本開示の様々な実施形態のこれらおよび他の特徴、態様、および利点は、以下の説明および添付の特許請求の範囲を参照すると、よりよく理解されよう。本明細書に組み込まれ、本明細書の一部を構成する添付の図面は、本開示の例示的実施形態を示し、この説明とともに、関連する原理について説明するために役立つ。

【００１５】

当業者を対象とする実施形態の詳細な説明が本明細書に記載され、本明細書は添付の図を参照する。

【図面の簡単な説明】

【００１６】

【図１】本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する例示的な機械学習済み言語モデルのブロック図である。

【図２】本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する例示的な機械学習済み言語モデルのブロック図である。

【図３】本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する機械学習済み言語モデルのための例示的な訓練プロセスのブロック図である。

【図４】本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する機械学習済み言語モデルのための例示的な訓練プロセスのブロック図で

ある。
【図５Ａ】本開示の例示的な実施形態による、例示的コンピューティングシステムのブロック図である。
【図５Ｂ】本開示の例示的な実施形態による、例示的コンピューティングデバイスのブロック図である。
【図５Ｃ】本開示の例示的な実施形態による、例示的コンピューティングデバイスのブロック図である。
【発明を実施するための形態】
【００１７】
複数の図にわたって繰り返される参照番号は、様々な実装形態において同じ特徴を識別するものである。
【００１８】
概要
一般に、本開示は、文脈テキスト生成のサービスにおいて中間テキスト分析(たとえば、APIなどの構造ツールの使用を含む)を生成する1つまたは複数の機械学習済み言語モデルを含むかつ/または活用するシステムおよびモデルを対象にする。たとえば、コンピューティングシステムは、1つまたは複数の文脈テキストトークンを含む文脈テキスト文字列を取得することができる。コンピューティングシステムは、1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列を生成するために、機械学習済み言語モデルを用いて文脈テキスト文字列を処理することができる。コンピューティングシステムは、1つまたは複数の出力テキストトークンを含む出力テキスト文字列を生成するために、機械学習済み言語モデルを用いて1つまたは複数の中間テキスト文字列を処理することができる。1つまたは複数の中間テキスト文字列は、出力テキスト文字列をサポートする文脈テキスト文字列のテキスト分析を含むことができる。
【００１９】
したがって、本開示の態様は、文脈テキスト入力に応答して出力テキストを生成する(たとえば、質問または前の対話に対する応答を生成する)前に(たとえば、そのサービス中に)テキスト分析を生成するようにモデルに教示することによって、知識、基礎訓練(grounding)、および機械学習済み言語モデルの解釈可能性を改善する。そのような中間テキスト分析の生成は、モデル出力の解釈可能性を改善することができる。特に、中間テキスト分析は、モデルが文脈入力に応答して出力をどのように生成したかを解釈または理解するために再検討または検査され得る。これはまた、特定のタスクをサービスすることにおいて出力の信頼性および/または適合性の評価を容易にし得る。
【００２０】
本開示の別の態様によれば、いくつかの実装形態では、テキスト分析は、追加の情報へのアクセスを提供する構造ツールの使用を含むかつ/または活用することができる。たとえば、中間テキスト分析に含まれる1つまたは複数の中間テキストトークンは、文脈テキスト文字列に含まれないおよび/またはモデルが訓練された訓練データに含まれない追加の情報にアクセスするために構造ツールの使用を引き起こす少なくとも1つのツールトークンを含むことができる。したがって、言語モデルは、最新式、事実、ドメイン固有、クライアントもしくはユーザ固有などであり得る追加の情報へのアクセスを有するためにそのような構造ツールを呼び出して使用することができる。これは、テキスト出力を形成するときに言語モデルに利用可能な知識を改善し、様々な使用事例に対する様々な情報源の導入を可能にすることによってシステムの柔軟性をさらに改善する。機械学習プロセスは、そのようなサービス、たとえばツールトークンの呼び出しにおける計算オーバーヘッドを最小化するために適用され得、それらが生成される順序は、レイテンシおよび/またはネットワーク使用などの計算オーバーヘッドを最小化するように適応され得るので、本開示の手法は、外部ツールとの改善されたまたは最適化された一体化を達成し得る。
【００２１】
例として、機械学習済み言語モデルの構造ツールは、データベースから追加の情報にア

クセスするためのデータベース検索、アプリケーションプログラミングインターフェース(API)を介して追加の情報を要求して受信するためのAPI呼び出し、入力テキストトークン上で1つまたは複数の動作のシーケンスを実行するプログラミング言語インタープリタ、検索エンジン、ナレッジグラフまたはデジタルアシスタントからの結果をクエリするクエリサービス、通信(たとえば、電子メール、ショートメッセージサービスメッセージ、マルチメディアメッセージングサービスメッセージ、アプリケーションベースのチャットメッセージなど)を生成して別のデバイスまたはユーザに送信する通信クライアント、および/または追加の情報へのアクセスを生成するかまたは場合によっては提供する様々な他の形式の構造ツールを含めるためのアクセスを有し得る。したがって、構造ツールは、情報を調べることに限定されず、副次効果を有すること、または行動(たとえば、会議を予約すること、何かを購買すること、チケットを人に提出することなど)を引き起こすこともできる。

【００２２】

本明細書で説明する機械学習済み言語モデルは、いくつかの異なる手法において訓練され得る。一例では、ヒューマンボランティアまたはクラウドワーカーは、いくつかの(たとえば、数千の)例に対して例示的な中間分析テキストを生成することができる。たとえば、ヒューマンワーカーは、文脈入力テキストと出力テキストのペアを与えられ、ヒューマンワーカーは、出力テキストをもたらすかまたは場合によってはサポートする文脈入力テキストの分析を示す中間分析テキストを生成することができる。ヒューマンワーカーは、構造ツールおよびそのようなツールのそれらの使用へのアクセスを与えられ得、取得された対応する情報は、例示的な中間分析テキスト内に含められ得る。

【００２３】

中間分析(事例を訓練中、またはモデルによって生成されるときのいずれか)は、いくつかの例では、代数の問題に対する多段階解法など、人間が読み取れる形成におけるステップバイステップ論理を含むことができる。それは、本明細書の他の場所で説明するように、データベース、pythonインタープリタ、検索エンジンなど、外部のテキストツーテキストツールの使用も含むことができる。いくつかの実装形態では、中間分析セクションにおけるツール使用は、どのツールが使用されたかを明記し、ツールの入力および出力を記述する特別なタグによってマークおよび/またはトリガされ得る。中間テキストは、ツール使用の複数のインスタンス、ならびに任意の量の自由形式テキストを含むことができる。

【００２４】

したがって、中間分析は、入力パラメータの構造化リストを取り込み、構造化出力(たとえば、do_thing(a: int, b: List[str]) --〉 response_object)を戻すツール(たとえば、API)の使用を含むことができる。しかしながら、任意の構造化入力/出力はまた、Google ProtosまたはJSONのテキストシリアライゼーションなど、いくつかのシリアライゼーション方法を使用して自由テキストにシリアライズされ、フリーテキストからパースされ得る。その観点から、テキストツーテキストインターフェースは、構造化インターフェースの上位集合であり得る。

【００２５】

例示的な訓練データセットを生成するために、ヒューマンアノテータによって生成された例示的な中間分析テキストが、訓練タプルを形成するために文脈入力テキストと出力テキストのペアと組み合わされ得、訓練タプルは、1つまたは複数の文脈テキストトークンを含む例示的な文脈テキスト文字列と、1つまたは複数の中間テキストトークンを含む1つまたは複数の例示的な中間テキスト文字列と、1つまたは複数の出力テキストトークンを含む例示的な出力テキストストリングとを含む。したがって、いくつかの実装形態では、ヒューマンアノテータは、例示的な文脈テキスト文字列および例示的な出力テキスト文字列を与えられ得、ヒューマンアノテータは、例示的な中間テキスト文字列を生成することができる。他の実装形態では、ヒューマンアノテータは、例示的な文脈テキスト文字列だけを与えられ得、ヒューマンアノテータは、例示的な中間テキスト文字列と例示的な出力テキスト文字列の両方を生成することができる。

【００２６】
本明細書で説明するような訓練データセットは、言語モデルを訓練するために使用され得る。たとえば、訓練データセットは、テラスケールの教師なしデータ上で事前訓練されたモデルを微調整するために使用され得る。一例として、モデルは、訓練タプルに含まれる次のトークン(たとえば、次の中間テキストトークンまたは次の出力テキストトークン)を予測するためにモデルを使用することによって訓練され得る。損失関数は、次のトークンを予測するためにモデルの能力を評価するために使用され得る。モデルのパラメータは、損失関数に基づいて(たとえば、逆伝搬ベースの技法を介して)更新され得る。いくつかの実装形態では、各訓練タプル上でモデルを訓練することは、中間テキスト文字列に含まれる各トークン上でトークンごとに反復して訓練し、続いて各トークンが出力テキスト文字列に含められることを含むことができる。
【００２７】
別の例では、言語モデルは、強化学習手法を使用して文脈言語生成のサービスにおいて中間テキストを生成するように訓練され得る。たとえば、モデルによって生成された中間テキストおよび/または出力テキストの態様は、報酬を決定するために目的関数によって評価され得、次いで報酬はモデルパラメータを更新するために使用され得る。
【００２８】
次いで、推論時に、言語モデルは、入力を与えられた中間分析を生成するために使用され得る。いくつかの実装形態では、中間分析は、一度に1つのトークンを生成され得る。いくつかの実装形態では、モデルが外部ツールへの入力の生成を終えるたびに、ツール自体が、ツール出力を生成するためにこの入力によって呼び出され、ツール出力は中間テキストに付加され、モデルはそこから生成することを継続する。
【００２９】
したがって、本開示の態様は、同じくテキストである中間分析部を有するために、(入力、出力)訓練例および言語生成パラダイムを拡張することを提案する。そのため、単に入力を与えられて出力を生成する代わりに、言語モデルは、入力を与えられて中間分析を生成し、次いで入力および中間分析を与えられて出力を生成するように学習することができる。
【００３０】
したがって、対話エージェント(または、任意の他の文脈テキスト生成モデル)が、与えられた文脈に対する応答を直接生成することができるために、(文脈、応答)ペア上で一般的に訓練されるのに反して、本開示の例示的な実装形態では、言語モデルは、代わりに、(文脈、中間分析、応答)トリプル上で訓練することができ、モデルは、(中間分析┆文脈)および(応答┆文脈、中間分析)を生成するように学習する。
【００３１】
いくつかの実装形態では、本明細書で説明するように生成された出力テキストは、オーディオ出力を生成するために音声合成システム(text to speech system)を使用してさらに処理され得る。別の例として、入力テキストが、テキストシステムに対する発話を使用してオーディオ入力から生成され得る。たとえば、仮想アシスタントが、オーディオ入力および出力を介してユーザと対話することができ、オーディオ/発話からテキストへの変換は、仮想アシスタントによる処理が本明細書で説明するテキストドメイン内で発生することを可能にするために使用され得る。
【００３２】
本開示のシステムおよび方法は、いくつかの技術的効果と利益とを提供する。一例として、提案されたモデルは、改善された解釈可能性を示す。たとえば、モデルによって生成された中間テキスト分析は、モデルが文脈入力に応答して出力をどのように生成したかを解釈または理解するために再検討または検査され得る。改善された解釈可能性は、プロセッサ使用量、メモリ使用量などの計算リソースのより効率的な使用につながることができる。たとえば、言語モデル出力における解釈可能性の欠如は、モデル出力に対する信頼または信用の欠如をもたらすことがあり、それは、モデルの出力の正確さまたは有用性を「ダブルチェック」しようとする、不必要なオーバーヘッドまたは他の努力(たとえば、コン

ピュータ化された動作)をもたらすことがある。解釈可能性を改善することによって、コンピュータ化されたシステムにおける信頼が改善され得る。特に、モデル出力の信頼性は、特定のタスクに対するシステムの有用性を確立するために検証および/または評価され得る。

【００３３】

別の例示的な技術的効果および利益として、提案された手法は、言語モデルが追加の事実情報などの追加の情報にアクセスするために構造ツールを活用することを可能にする。したがって、言語モデルは、最新式、事実、ドメイン固有、クライアントもしくはユーザ固有などであり得る追加の情報へのアクセスを有するためにそのような構造ツールを呼び出して使用することができる。これは、テキスト出力を形成するときに言語モデルに利用可能な知識を改善し、様々な使用事例に対する様々な情報源の導入を可能にすることによってシステムの柔軟性をさらに改善する。

【００３４】

モデルの出力の品質を改善することに加えて、構造ツールの提案される使用はまた、プロセッサ使用量、メモリ使用量、ネットワーク帯域幅などの計算リソースの節約につながる。具体的には、以前の言語モデルに利用可能な知識は、それらの訓練データセットに含まれる(およびそれらから学習される)情報および/または文脈テキスト入力の中で導入された情報に制約された。したがって、事実情報のそれらの知識は極めて限定され、一般に時間的に凍結されている。そのため、事実情報を含む出力を作成することを要求されるとき、モデルは、一般的に、間違った事実を錯覚するか、または古い情報を供給するかのいずれかである。それゆえ、言語モデル全体は、実世界の事実の変化につれて言語モデルを最新のものに保つため、言語モデルをユーザ情報の新しいドメインまたはセットに移植するため、または場合によってはモデルを新しい情報が未解決であった新しい状況に展開するために、再訓練される必要がある。言語モデルの再訓練は、プロセッサ使用量、メモリ使用量、ネットワーク帯域幅などの計算リソースの使用を必要とする。

【００３５】

しかしながら、本開示によって提案される構造ツールの使用は、実世界の事実の変化につれて言語モデルを最新のものに保つため、言語モデルをユーザ情報の新しいドメインまたはセットに移植するため、または場合によってはモデルを新しい情報が未解決であった新しい状況に展開するためにモデルを再訓練する必要を取り除く。代わりに、モデルは、単に、最新式、事実、ドメイン固有、クライアントもしくはユーザ固有などであり得る追加の情報へのアクセスを(たとえば、構造ツールを介して)を与えられ得る。したがって、モデルは、異なるドメイン、使用、ユーザなどに容易に移植され得、かつ/またはモデルを再訓練する必要なしに、最新の事実情報を活用する応答を提供することができ、それにより計算リソースを大幅に節約する。(潜在的に外部の)情報源とインターフェースし得る中間分析の形式の文脈を符号化することによって、処理は、情報および/または機能の提供における技術的制約の解決に寄与し得る。

【００３６】

同様に、別の例示的な技術的効果が、様々な入力に応答するために必要な情報のすべてを(たとえば、学習済みの関係性の形式で)記憶することを必要とせずに、情報を取得するために外部ソースを活用するためにモデルの能力から導出される。特に、過去の手法は、様々な入力および出力の間の関係性を学習して記憶するために十分なサイズ(たとえば、パラメータの数)を有する大きいモデルの記憶および(たとえば、制約されたメモリおよび/またはバッテリの利用可能性を有するユーザデバイス上での)使用を必要とした。対照的に、本開示のいくつかの例示的な実装形態は、ユーザデバイスまたは他のモバイルクライアントもしくはブラウザ上で存在するために「薄い」(より小さい)モデルを可能にし得る。薄いモデルは、バッテリ、計算、記憶、更新などを節約するために様々な構造ツール(たとえば、クラウドサービス)を活用することができる。したがって、構造ツールへのアクセスを有するより小さいモデルが、大きい自己完結型のモデルと同様の、またはそれより優れた性能を達成することができ、それによりメモリ使用量、ネットワーク帯域幅、エネルギー

消費などの計算リソースを節約する。

【００３７】

次に図を参照しながら、本開示の例示的実施形態について、さらに詳細に説明する。

【００３８】

中間テキスト分析を生成する例示的な言語モデル

図1は、本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する例示的な機械学習済み言語モデルのブロック図を示す。具体的には、言語モデル14は、1つまたは複数の文脈テキストトークンを含む文脈テキスト文字列12を受信することができる。言語モデル14は、1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列16を生成するために、文脈テキスト文字列12を処理することができる。言語モデル14は、1つまたは複数の出力テキストトークン18を含む出力テキスト文字列を生成するために、1つまたは複数の中間テキスト文字列16を処理することができる。1つまたは複数の中間テキスト文字列16は、出力テキスト文字列18を生成するために文脈テキスト文字列12の論理分析をサポートする、もたらす、証明する、または場合によっては示す文脈テキスト文字列12のテキスト分析を含むことができる。

【００３９】

本開示の態様によれば、いくつかの実装形態では、1つまたは複数の中間テキストトークン16は、文脈テキスト文字列12に含まれない追加の情報にアクセスするために構造ツール15の使用を引き起こす少なくとも1つのツールトークンを含むことができる。いくつかの実装形態では、構造ツール15は、データベースから追加の情報にアクセスするためのデータベース検索を含むことができる。いくつかの実装形態では、構造ツール15は、アプリケーションプログラミングインターフェース(API)を介して追加の情報を要求して受信するためのAPI呼び出しを含むことができる。いくつかの実装形態では、構造ツール15は、入力テキストトークン上で1つまたは複数の動作のシーケンスを実行するプログラミング言語インタープリタを含むことができる。いくつかの実装形態では、構造ツール15は、検索エンジン、ナレッジグラフ、またはデジタルアシスタントからの結果をクエリするクエリサービスを含むことができる。ツールトークンに加えて、1つまたは複数の中間テキストトークン16は、少なくとも1つの自然言語テキストトークンをさらに含むことができる。

【００４０】

いくつかの実装形態では、機械学習済み言語モデル14がツールトークンを生成するとき、コンピューティングシステムは、機械学習済み言語モデルを中断することと、追加の情報にアクセスするために構造ツール15を実行することと、1つまたは複数の中間テキスト文字列16の現在のバージョンに追加の情報を付加することと、1つまたは複数の中間テキスト文字列16の現在のバージョンおよび付加された追加の情報に基づいて機械学習済み言語モデル14を用いてテキスト生成を再開することとを行うことができる。

【００４１】

いくつかの実装形態では、機械学習済み言語モデル14は、トークンごとに動作する。そのような実装形態の一部では、1つまたは複数の中間テキスト文字列16を生成するとき、言語モデル14は、各生成された中間テキストトークン16を再帰的に入力として受信する。

【００４２】

したがって、いくつかの実装形態では、1つまたは複数の中間テキストトークンを含む1つまたは複数の中間テキスト文字列16を生成するために機械学習済み言語モデル14を用いて文脈テキスト文字列12を処理することが、いくつかの反復にわたって実行され得る。第1の反復において、コンピューティングシステムは、1つまたは複数の中間テキストトークンを含む第1の中間テキスト文字列16を生成するために、機械学習済み言語モデル14を用いて文脈テキスト文字列12を処理することができる。次いで、コンピューティングシステムは、更新された文脈テキスト文字列を生成するために、文脈テキスト文字列12に第1の中間テキスト文字列16を付加することができる。次いで、1つまたは複数の追加の反復の各々に対しておよび機械学習済み言語モデルが閉止トークンを出力するまで、コンピュ

ーティングシステムは、1つまたは複数の中間テキストトークンを含む追加の中間テキスト文字列16を生成するために、機械学習済み言語モデル14を用いて更新された文脈テキスト文字列を処理することができる。コンピューティングシステムは、次の反復に対して更新された文脈テキスト文字列を生成するために、更新された文脈テキスト文字列に追加の中間テキスト文字列を付加することができる。

【００４３】

機械学習済み言語モデル14は、例として回帰型ニューラルネットワーク、マルチヘッドセルフアテンションモデル、シーケンスツーシーケンスモデル、および/または他の形式の言語モデルを含む様々なタイプのモデルであり得るか、またはそれらを含むことができる。言語モデルは、クローズモデル(cloze model)であり得るか、または左から右へのモデルであり得る。言語モデルは、随意に、エンコーダデコーダアーキテクチャを有することができる。

【００４４】

いくつかの例示的な実装形態では、機械学習済み言語モデル14は、質問応答モデルであり得、文脈テキスト文字列12は、質問であり得るか、またはそれを含むことができる。いくつかの例示的な実装形態では、機械学習済み言語モデル14は対話モデルであり得、文脈テキスト文字列12は対話履歴であり得るか、またはそれを含むことができる。

【００４５】

いくつかの実装形態では、文脈テキスト文字列12の少なくとも一部は、ユーザによって入力されたテキストを含むか、またはそれに対応する。いくつかの実装形態では、コンピューティングシステムは、ユーザに表示するために少なくとも出力テキスト文字列18を提供することができる。

【００４６】

図2は、本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する例示的な機械学習済み言語モデル14のブロック図を示す。特に、図2において文脈テキスト文字列12がベース言語モデル202に追加で入力されることを除いて、図2は図1と同様である。ベース言語モデル202は、いくつかの実装形態では、中間テキスト文字列を生成することなく、文脈テキスト文字列12からベース出力204を直接生成するように構成される。図2に示すように、ベース出力204は、文脈テキスト文字列12と組み合わされ(たとえば、それに付加または連結され)、組み合わされたストリングは、次いで機械学習済み言語モデル14に入力される。

【００４７】

この方式でのベース言語モデル202の使用は、機械学習済み言語モデル14の役割が、誤り訂正または「事実確認」の役割に変わることを可能にすることができる。特に、図1では、モデル14は主に、出力テキスト18を生成する責任を負う。対照的に、図2では、モデル14の役割は、ベース出力204に含まれる事実を補足または訂正することであり得る。このようにして、既存のベース言語モデル202は、構造ツール15へのアクセスを有するモデル14の追加を通して拡張または活用され得る。

【００４８】

たとえば、これは、様々なタスク、文脈、ドメイン、ユーザ、アプリケーションなどに対してすでに訓練されている、任意の数の様々な既存のベースモデルにモデル14およびツール15を適用することを可能にする。したがって、カスタム言語モデルをすでに有する任意のアプリケーションが、文脈言語出力を生成するときに事実情報または最新の情報の改善された使用を提供するために、追加のモデル14と組み合わされ得る。

【００４９】

図3は、本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する機械学習済み言語モデルのための例示的な訓練プロセスのブロック図を示す。具体的には、図3は、教師あり訓練手法を示す。

【００５０】

図3に示すように、いくつかの訓練テキストトークン312が取得され得る。訓練テキス

トトークン312の一部が、言語モデル314に入力され得る。モデル314は、訓練テキストトークン312に対する次の予測されるテキストトークン316を予測することができる。たとえば、次の予測されるテキストトークンは、例示的な中間テキストトークンであり得るか、または例示的な出力トークンであり得る。次の予測されるテキストトークン316は、損失関数318を使用して訓練テキストトークン312に含まれるグラウンドトゥルーステキストトークンと比較され得る。モデル314のパラメータは、(たとえば、対数損失関数または同様の)損失関数318に基づいて更新され得る。図3に示すプロセスは、訓練テキストトークン312に含まれる各テキストトークンに対して反復して連続的に実行され得る。たとえば、これを行うための1つの方法は、トークンシーケンス[〈tokenized-context〉, EOS, 〈tokenized-reasoning〉, EOS, 〈tokenized-response〉, EOS]上で左から右への言語モデルを訓練することである。プロセスは、いくつかの異なる訓練例にわたって実行され得る。

【００５１】

いくつかの例では、訓練データは、ボランティア/クラウドワーカーインターフェースを介して収集され得る。一例として、ヒューマンアノテータは、ベース言語モデルと対話することができる。ベース言語モデルが応答を発行した後、ヒューマンアノテータは、ユーザが「中間分析」に入ること、および/またはベース言語モデルの応答の出力を編集することを可能にするフィードバックインターフェースを開き得る。「中間分析」は、ベース言語ツールに対する呼び出しで開始することができ、ベース言語ツールの出力は、ベース言語モデルによって生成される、現在の文脈に対するいくつかの例示的な応答を含むことができる。次いで、ヒューマンアノテータは、追加のツール使用を含み得る中間分析を追加し得る。ツール使用を促進するために、ヒューマンアノテータがツールをクエリすることを可能にする形式が提供され得、ツールは、ツール入力/出力を中間分析に付加するためのボタンを有する。ヒューマンアノテータが中間分析を追加することを終了すると、ヒューマンアノテータは、随意に、ベースモデルの応答を修正して「セーブ」をクリックし、それによりヒューマンアノテータは会話インターフェースに戻る。変更はベースモデルの応答に反映され、何かを言うのは、ヒューマンアノテータの番である。いくつかの実装形態では、ベースモデルは、単に、使用された別のツールと見なされ得る。

【００５２】

図4は、本開示の例示的な実施形態による、文脈テキスト生成のサービスにおいてテキスト分析を生成する機械学習済み言語モデルのための例示的な訓練プロセスのブロック図を示す。具体的には、図4は、強化訓練手法を示す。

【００５３】

図4に示すように、1つまたは複数の文脈トークン412は、図1および図2で説明したような機械学習済み言語モデル414に入力され得る。モデル414は、中間テキスト文字列(たとえば、それは構造ツール415にアクセスすることまたはそれを活用することを含み得る)を生成することができる。最後に、モデル414は、1つの出力トークン416を生成することができる。報酬関数418は、出力トークン416に基づいて報酬を決定することができる。報酬関数418は、出力トークン416がどれほど十分に、いくつかの目的の満足度(たとえば、ユーザ満足度)を満たしたか、またはもたらしたかを決定することができる。随意に、出力トークン416は、ユーザまたは他の対話エージェントに供給され、追加の新しい文脈トークンをもたらすことができ、その文脈トークンは説明したプロセスを再開することができる。

【００５４】

中間分析による例示的なテキスト生成

このセクションは、様々な例示的なアプリケーションを示す中間テキスト分析のいくつかの例を提供する。

【００５５】

ベースエージェントを呼び出し、次いで事実確認/補足:

【００５６】

文脈:
【００５７】
エージェント: ハイ
【００５８】
ユーザ: ハイ、メキシコについて私に何か教えてもらえますか？
【００５９】
中間分析:
【００６０】
〈tool: base-pass-context〉 #が文脈を自動的に得る
【００６１】
〈tool-output〉
【００６２】
- メキシコは、公式にはメキシコ合衆国であり、北アメリカの南部にある国家です。
【００６３】
- メキシコは北でアメリカ合衆国に隣接し、南および西で太平洋に、南東でグアテマラ、ベリーズおよびカリブ海に、東でメキシコ湾に隣接しています。
【００６４】
- メキシコの大統領は国家の長であり、メキシコ政府の長でもあります。
【００６５】
〈/tool〉
【００６６】
〈tool: assistant〉
【００６７】
メキシコの現在の大統領
【００６８】
〈tool-output〉
【００６９】
Andres Manuel Lopez Obrador
【００７０】
〈/tool〉
【００７１】
応答:
【００７２】
- メキシコは、公式にはメキシコ合衆国であり、北アメリカの南部にある国家です。メキシコの現在の大統領は、Andres Manuel Lopez Obradorです。
【００７３】
エンドツーエンド顧客サポート
【００７４】
文脈1:
【００７５】
顧客: ハイ。パスワードがわからなくなってしまいました
【００７６】
中間分析1:
【００７７】
〈tool internet provider API〉
【００７８】
パスワードを忘れたときの対処法
【００７９】
〈tool-output〉
【００８０】

internetprovider.com/passwordに進んでください
【００８１】
〈/tool〉
【００８２】
出力1:
【００８３】
エージェント: ハイ。こんなことになってお気の毒です。あなたは
【００８４】
パスワードをinternetprovider.com/passwordにおいてリセットすることができます
【００８５】
文脈2:
【００８６】
顧客: オーケー...上手くいきました。しかしインターネットが遅いです
【００８７】
中間分析2:
【００８８】
〈tool internet provider API〉
【００８９】
インターネットが遅いときの対処方法
【００９０】
〈tool-output〉
【００９１】
カスタムインターネットプロバイダメーターを用いて速度を測定してください
【００９２】
〈/tool〉
【００９３】
〈tool meter〉
【００９４】
顧客id 777
【００９５】
〈tool-output〉
【００９６】
1234Mbps
【００９７】
〈/tool〉
【００９８】
〈tool cust-acct〉
【００９９】
顧客id777
【０１００】
〈tool-output〉
【０１０１】
John Doe
【０１０２】
インターネットプロバイダ最終プラン
【０１０３】
予期される速度: 1000Mbps
【０１０４】
〈/tool〉
【０１０５】

出力2:
【０１０６】
エージェント: オーケー。あなたが1000Mbpsパッケージを購入し、1234Mbpsをお持ちであることを確認しました。すべて問題なく進んでいます。更新を行いますか？
【０１０７】
例示的なデバイスおよびシステム
図5Aは、本開示の例示的実施形態による例示的コンピューティングシステム100のブロック図を示す。システム100は、ネットワーク180を介して通信可能に結合されている、ユーザコンピューティングデバイス102、サーバコンピューティングシステム130、およびトレーニング用コンピューティングシステム150を含む。
【０１０８】
ユーザコンピューティングデバイス102は、たとえば、パーソナルコンピューティングデバイス(たとえば、ラップトップまたはデスクトップ)、モバイルコンピューティングデバイス(たとえば、スマートフォンまたはタブレット)、ゲーミングコンソールもしくはコントローラ、ウェアラブルコンピューティングデバイス、埋込み型コンピューティングデバイス、または任意の他のタイプのコンピューティングデバイスなどの、任意のタイプのコンピューティングデバイスであってもよい。
【０１０９】
ユーザコンピューティングデバイス102は、1つまたは複数のプロセッサ112およびメモリ114を含む。1つまたは複数のプロセッサ112は、どの適切な処理デバイス(たとえば、プロセッサコア、マイクロプロセッサ、ASIC、FPGA、コントローラ、マイクロコントローラなど)であってもよく、1つのプロセッサまたは動作可能に接続されている複数のプロセッサであってよい。メモリ114は、RAM、ROM、EEPROM、EPROM、フラッシュメモリデバイス、磁気ディスクなど、およびそれらの組合せのような、1つまたは複数の非一時的コンピュータ可読記憶媒体を含み得る。メモリ114は、データ116と、ユーザコンピューティングデバイス102に動作を実施させるようにプロセッサ112によって実行される命令118とを記憶することができる。
【０１１０】
いくつかの実装形態では、ユーザコンピューティングシステム102は、1つまたは複数の機械学習済みモデル120を記憶するか、または含むことができる。たとえば、機械学習済みモデル120は、ニューラルネットワーク(たとえば、ディープニューラルネットワーク)または非線形モデルおよび/もしくは線形モデルを含む他のタイプの機械学習済みモデルなど、様々な機械学習済みモデルであってよく、またはそうでなければ、それらの機械学習済みモデルを含むことができる。ニューラルネットワークは、フィードフォワードニューラルネットワーク、回帰型ニューラルネットワーク(たとえば、長短期メモリ回帰型ニューラルネットワーク)、畳み込みニューラルネットワーク、または他の形のニューラルネットワークを含み得る。いくつかの例示的な機械学習済みモデルは、セルフアテンションなどのアテンションメカニズムを活用することができる。たとえば、いくつかの例示的な機械学習済みモデルは、マルチヘッドセルフアテンションモデル(たとえば、トランスフォーマモデル)を含むことができる。例示的な機械学習済みモデル120については、図1～図4を参照して論じる。
【０１１１】
いくつかの実装形態では、1つまたは複数の機械学習済みモデル120は、ネットワーク180を介してサーバコンピューティングシステム130から受信され、ユーザコンピューティングデバイスメモリ114に記憶され、次いで、1つまたは複数のプロセッサ112によって使われ、またはそうでなければ実装され得る。いくつかの実装形態では、ユーザコンピューティングデバイス102は、(たとえば、言語生成タスクの複数の例にわたってパラレル言語生成を実行するために)単一機械学習済みモデル120の複数の並列インスタンスを実装することができる。
【０１１２】

追加または代替として、1つまたは複数の機械学習済みモデル140は、クライアント-サーバ関係に従ってユーザコンピューティングデバイス102と通信するサーバコンピューティングシステム130に含まれ、またはそうでなければ、サーバコンピューティングシステム130によって記憶され、実装され得る。たとえば、機械学習済みモデル140は、ウェブサービス(たとえば、質問応答サービスなどの言語生成サービス、対話サービス(たとえば、「チャットボット」もしくはデジタルアシスタントによって使用される)など)の一部としてサーバコンピューティングシステム130によって実装され得る。したがって、1つまたは複数のモデル120が、ユーザコンピューティングデバイス102において記憶され、実装されてよく、かつ/または1つもしくは複数のモデル140が、サーバコンピューティングシステム130において記憶され、実装されてよい。モデル120および/または140は、質問応答サービス、対話サービス(たとえば、「チャットボット」もしくはデジタルアシスタントによって使用される)などの任意の言語生成サービスによって使用され得る。

【０１１３】

ユーザコンピューティングデバイス102は、ユーザ入力を受信する1つまたは複数のユーザ入力構成要素122も含み得る。たとえば、ユーザ入力構成要素122は、ユーザ入力オブジェクト(たとえば、指またはスタイラス)のタッチに敏感な、タッチ感応構成要素(たとえば、タッチ感応表示画面またはタッチパッド)であってよい。タッチ感応構成要素は、仮想キーボードを実装するのに役立ち得る。他の例示的ユーザ入力構成要素は、マイクロフォン、従来のキーボード、またはユーザがユーザ入力を与えることができる他の手段を含む。

【０１１４】

サーバコンピューティングシステム130は、1つまたは複数のプロセッサ132およびメモリ134を含む。1つまたは複数のプロセッサ132は、どの適切な処理デバイス(たとえば、プロセッサコア、マイクロプロセッサ、ASIC、FPGA、コントローラ、マイクロコントローラなど)であってもよく、1つのプロセッサまたは動作可能に接続されている複数のプロセッサであってよい。メモリ134は、RAM、ROM、EEPROM、EPROM、フラッシュメモリデバイス、磁気ディスクなど、およびそれらの組合せのような、1つまたは複数の非一時的コンピュータ可読記憶媒体を含み得る。メモリ134は、データ136と、サーバコンピューティングシステム130に動作を実施させるようにプロセッサ132によって実行される命令138とを記憶することができる。

【０１１５】

いくつかの実装形態では、サーバコンピューティングシステム130は、1つまたは複数のサーバコンピューティングデバイスを含むか、またはそうでなければ、1つまたは複数のサーバコンピューティングデバイスによって実装される。サーバコンピューティングシステム130が複数のサーバコンピューティングデバイスを含む事例では、そのようなサーバコンピューティングデバイスは、順次コンピューティングアーキテクチャ、並列コンピューティングアーキテクチャ、またはそれらの何らかの組合せに従って動作することができる。

【０１１６】

上述したように、サーバコンピューティングシステム130は、1つまたは複数の機械学習済みモデル140を記憶するか、またはそうでなければ含むことができる。たとえば、モデル140は、様々な機械学習済みモデルであってよく、または、そうでなければそれらを含んでよい。例示的機械学習済みモデルは、ニューラルネットワークまたは他のマルチレイヤ非線形モデルを含む。例示的ニューラルネットワークは、フィードフォワードニューラルネットワーク、ディープニューラルネットワーク、回帰型ニューラルネットワーク、および畳み込みニューラルネットワークを含む。いくつかの例示的な機械学習済みモデルは、セルフアテンションなどのアテンションメカニズムを活用することができる。たとえば、いくつかの例示的な機械学習済みモデルは、マルチヘッドセルフアテンションモデル(たとえば、トランスフォーマモデル)を含むことができる。例示的モデル140については、図1～図4を参照して論じる。

【０１１７】
ユーザコンピューティングデバイス102および/またはサーバコンピューティングシステム130は、ネットワーク180を介して通信可能に結合されるトレーニング用コンピューティングシステム150との対話により、モデル120および/または140をトレーニングすることができる。トレーニング用コンピューティングシステム150は、サーバコンピューティングシステム130とは別個であってよく、またはサーバコンピューティングシステム130の一部分であってよい。
【０１１８】
トレーニング用コンピューティングシステム150は、1つまたは複数のプロセッサ152およびメモリ154を含む。1つまたは複数のプロセッサ152は、どの適切な処理デバイス(たとえば、プロセッサコア、マイクロプロセッサ、ASIC、FPGA、コントローラ、マイクロコントローラなど)であってもよく、1つのプロセッサまたは動作可能に接続されている複数のプロセッサであってよい。メモリ154は、RAM、ROM、EEPROM、EPROM、フラッシュメモリデバイス、磁気ディスクなど、およびそれらの組合せのような、1つまたは複数の非一時的コンピュータ可読記憶媒体を含み得る。メモリ154は、データ156と、トレーニング用コンピューティングシステム150に動作を実施させるようにプロセッサ152によって実行される命令158とを記憶することができる。いくつかの実装形態では、トレーニング用コンピューティングシステム150は、1つまたは複数のサーバコンピューティングデバイスを含むか、またはそうでなければ、サーバコンピューティングデバイスによって実装される。
【０１１９】
トレーニング用コンピューティングシステム150は、ユーザコンピューティングデバイス102および/またはサーバコンピューティングシステム130において記憶された機械学習済みモデル120および/または140を、たとえば、誤差逆伝播など、様々なトレーニングまたは学習技法を使ってトレーニングするモデル訓練器160を含み得る。たとえば、損失関数は、(たとえば、損失関数の勾配に基づいて)モデルの1つまたは複数のパラメータを更新するために、モデルを通して逆伝搬され得る。平均2乗誤差、尤度損失、交差エントロピー損失、ヒンジ損失、および/または様々な他の損失関数など、様々な損失関数が使用され得る。勾配降下技法は、いくつかのトレーニング反復に対してパラメータを反復的に更新するために使用され得る。
【０１２０】
いくつかの実装形態では、誤差逆伝播を実行することは、短縮された時通的逆伝播を実行することを含み得る。モデル訓練器160は、トレーニングされるモデルの汎化能力を向上するために、いくつかの汎化技法(たとえば、重み減衰、ドロップアウトなど)を実施することができる。
【０１２１】
特に、モデル訓練器160は、トレーニングデータのセット162に基づいて、機械学習済みモデル120および/または140をトレーニングすることができる。訓練データ162は、たとえば、複数の訓練タプルを含むことができる。各訓練タプルは、1つまたは複数の文脈テキストトークンを含む例示的な文脈テキスト文字列と、1つまたは複数の中間テキストトークンを含む1つまたは複数の例示的な中間テキスト文字列と、1つまたは複数の出力テキストトークンを含む例示的な出力テキスト文字列とを含むことができる。
【０１２２】
いくつかの実装形態では、ユーザが承諾を与えた場合、トレーニング例は、ユーザコンピューティングデバイス102によって提供され得る。したがって、そのような実装形態では、ユーザコンピューティングデバイス102に提供されるモデル120は、ユーザコンピューティングデバイス102から受信されるユーザ固有データに対してトレーニングコンピューティングシステム150によってトレーニングされ得る。場合によっては、このプロセスは、モデルの個人化と呼ばれることがある。
【０１２３】

モデル訓練器160は、所望の機能性を提供するのに使用されるコンピュータ論理を含む。モデル訓練器160は、汎用プロセッサを制御するハードウェア、ファームウェア、および/またはソフトウェアで実装することができる。たとえば、いくつかの実装形態では、モデル訓練器160は、記憶デバイス上に記憶され、メモリにロードされ、1つまたは複数のプロセッサによって実行されるプログラムファイルを含む。他の実装形態では、モデル訓練器160は、RAM、ハードディスク、または光学もしくは磁気媒体などの有形コンピュータ可読記憶媒体に記憶されるコンピュータ実行可能命令の1つまたは複数のセットを含む。

【０１２４】

ネットワーク180は、ローカルエリアネットワーク(たとえば、イントラネット)、ワイドエリアネットワーク(たとえば、インターネット)、またはそれらの何らかの組合せなど、どのタイプの通信ネットワークであってもよく、任意の数のワイヤードまたはワイヤレスリンクを含み得る。概して、ネットワーク180を介した通信は、非常に様々な通信プロトコル(たとえば、TCP/IP、HTTP、SMTP、FTP)、符号化もしくはフォーマット(たとえば、HTML、XML)、および/または保護方式(たとえば、VPN、セキュアHTTP、SSL)を使って、どのタイプのワイヤードおよび/またはワイヤレス接続を介しても搬送することができる。

【０１２５】

図5Aは、本開示を実装するために使用され得る1つの例示的コンピューティングシステムを示す。他のコンピューティングシステムが同様に使用されてもよい。たとえば、いくつかの実装形態では、ユーザコンピューティングデバイス102は、モデル訓練器160および訓練データセット162を含み得る。そのような実装形態では、モデル120は、ユーザコンピューティングデバイス102においてローカルにトレーニングされることと使われることの両方が可能である。そのような実装形態のいくつかでは、ユーザコンピューティングデバイス102は、ユーザ固有データに基づいて、モデル訓練器160を実装して、モデル120を個人化し得る。

【０１２６】

図5Bは、本開示の例示的実施形態に従って実行する例示的コンピューティングデバイス10のブロック図を示す。コンピューティングデバイス10は、ユーザコンピューティングデバイスまたはサーバコンピューティングデバイスであってよい。

【０１２７】

コンピューティングデバイス10は、いくつかのアプリケーション(たとえば、アプリケーション1～N)を含む。各アプリケーションは、それ自体の機械学習ライブラリおよび機械学習済みモデルを含む。たとえば、各アプリケーションは、機械学習済みモデルを含み得る。例示的アプリケーションは、テキストメッセージングアプリケーション、eメールアプリケーション、ディクテーションアプリケーション、仮想キーボードアプリケーション、ブラウザアプリケーションなどを含む。

【０１２８】

図5Bに示すように、各アプリケーションは、コンピューティングデバイスのいくつかの他の構成要素、たとえば、1つもしくは複数のセンサ、コンテキストマネージャ、デバイス状態構成要素、および/または追加構成要素などと通信することができる。いくつかの実装形態では、各アプリケーションは、API(たとえば、パブリックAPI)を使って、各デバイス構成要素と通信することができる。いくつかの実装形態では、各アプリケーションによって使用されるAPIは、そのアプリケーションに固有である。

【０１２９】

図5Cは、本開示の例示的実施形態に従って実行する例示的コンピューティングデバイス50のブロック図を示す。コンピューティングデバイス50は、ユーザコンピューティングデバイスまたはサーバコンピューティングデバイスであってよい。

【０１３０】

コンピューティングデバイス50は、いくつかのアプリケーション(たとえば、アプリケーション1～N)を含む。各アプリケーションは、中央インテリジェンスレイヤと通信する

。例示的アプリケーションは、テキストメッセージングアプリケーション、eメールアプリケーション、ディクテーションアプリケーション、仮想キーボードアプリケーション、ブラウザアプリケーションなどを含む。いくつかの実装形態では、各アプリケーションは、API(たとえば、すべてのアプリケーションにわたる共通API)を使って、中央インテリジェンスレイヤ(およびその中に記憶されるモデル)と通信することができる。

【０１３１】

中央インテリジェンスレイヤは、いくつかの機械学習済みモデルを含む。たとえば、図5Cに示すように、それぞれの機械学習済みモデルが、各アプリケーションに与えられ、中央インテリジェンスレイヤによって管理され得る。他の実装形態では、2つ以上のアプリケーションが、単一の機械学習済みモデルを共有することができる。たとえば、いくつかの実装形態では、中央インテリジェンスレイヤは、アプリケーションすべてに単一モデルを提供することができる。いくつかの実装形態では、中央インテリジェンスレイヤは、コンピューティングデバイス50のオペレーティングシステムに含まれるか、またはそうでなければ、オペレーティングシステムによって実装される。

【０１３２】

中央インテリジェンスレイヤは、中央デバイスデータレイヤと通信することができる。中央デバイスデータレイヤは、コンピューティングデバイス50向けのデータの集中型リポジトリであってよい。図5Cに示すように、中央デバイスデータレイヤは、コンピューティングデバイスのいくつかの他の構成要素、たとえば、1つまたは複数のセンサ、コンテキストマネージャ、デバイス状態構成要素、および/または追加構成要素などと通信することができる。いくつかの実装形態では、中央デバイスデータレイヤは、API(たとえば、プライベートAPI)を使用して、各デバイス構成要素と通信することができる。

【０１３３】

追加開示

本明細書において論じた技術は、サーバ、データベース、ソフトウェアアプリケーション、および他のコンピュータベースのシステム、ならびに行われるアクションおよびそのようなシステムとの間で送られる情報を参照する。コンピュータベースのシステムに固有の柔軟性によって、構成要素の間のタスクおよび機能の多種多様な可能な構成、組合せ、および分割が可能になる。たとえば、本明細書で説明するプロセスは、単一のデバイスもしくは構成要素、または組合せにおいて働く複数のデバイスもしくは構成要素を使用して実装され得る。データベースおよびアプリケーションは、単一のシステム上で実装されるか、または複数のシステムにわたって分散され得る。分散構成要素は、順次または並行して動作することができる。

【０１３４】

本主題は、その様々な特定の例示的実施形態に関して詳細に説明されてきたが、各例は、本開示の限定ではなく、説明として与えられる。当業者は、上記を理解すると、そのような実施形態の改変、変形、および等価物を容易に作り出すことができる。したがって、本開示は、当業者には容易に明らかであろうように、本主題へのそのような修正、変形および/または追加を含めることを排除しない。たとえば、1つの実施形態の一部として示されるかまたは説明される特徴は、またさらなる実施形態をもたらすために、別の実施形態とともに使用され得る。したがって、本開示がそのような改変、変形、および等価物をカバーすることが意図されている。

【符号の説明】

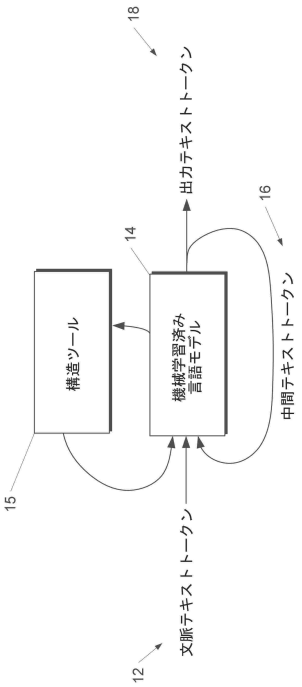
【０１３５】

10 コンピューティングデバイス
12 文脈テキスト文字列
14 言語モデル
15 構造ツール
16 中間テキスト文字列
18 出力テキスト文字列

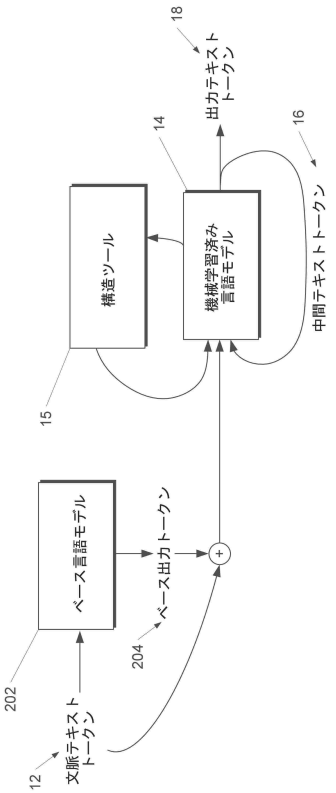
50 コンピューティングデバイス
100 コンピューティングシステム
102 ユーザコンピューティングデバイス
112 プロセッサ
114 メモリ
116 データ
118 命令
120 機械学習済みモデル
122 ユーザ入力構成要素
130 サーバコンピューティングシステム
132 プロセッサ
134 メモリ
136 データ
138 命令
140 機械学習済みモデル
150 トレーニング用コンピューティングシステム
152 プロセッサ
154 メモリ
156 データ
158 命令
160 モデル訓練器
162 トレーニングデータのセット
180 ネットワーク
202 ベース言語モデル
204 ベース出力
312 訓練テキストトークン
314 言語モデル
316 次の予測されるテキストトークン
318 損失関数
412 文脈トークン
414 機械学習済み言語モデル
415 構造ツール
416 出力トークン
418 報酬関数

【図面】

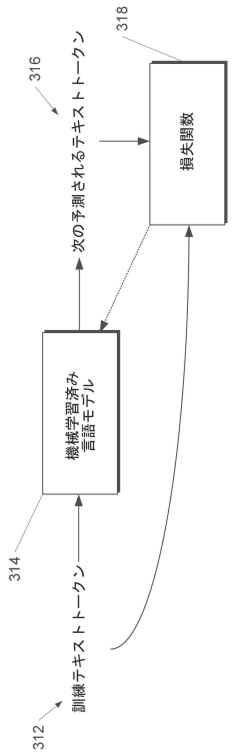
【図１】


18
出力テキストトークン
16
14
構造ツール
機械学習済み
言語モデル
中間テキストトークン
15
文脈テキストトークン
12

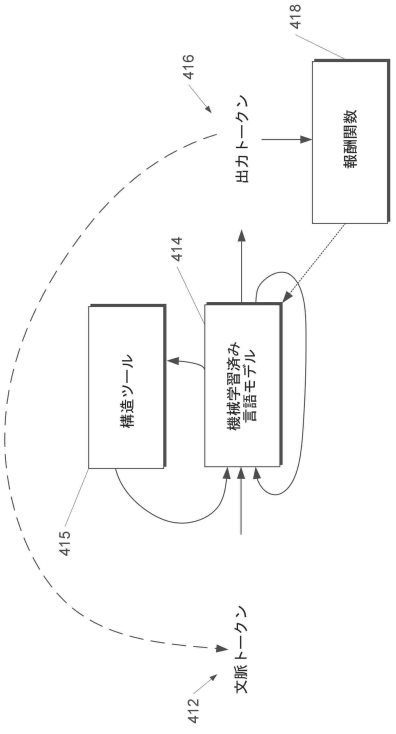
【図２】


18
出力テキスト
トークン
14
16
構造ツール
機械学習済み
言語モデル
中間テキストトークン
15
ベース言語モデル
ベース出力トークン
202
204
文脈テキスト
トークン
12

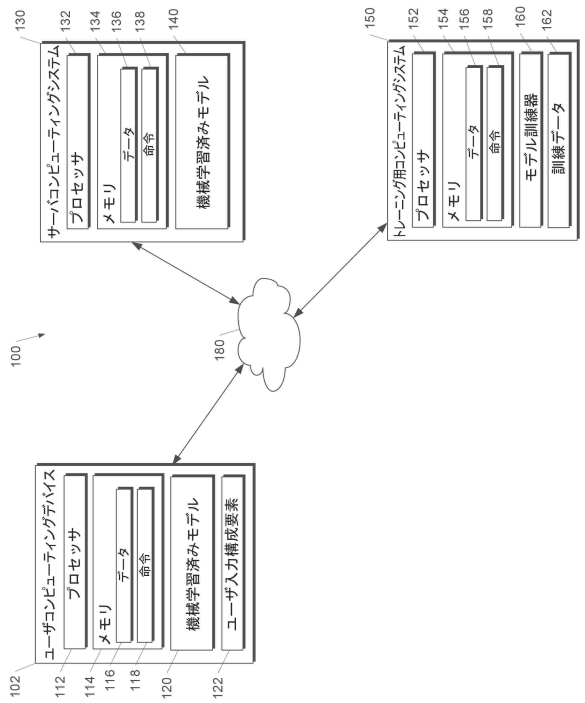
【図３】


318
316
次の予測されるテキストトークン
損失関数
機械学習済み
言語モデル
314
訓練テキストトークン
312

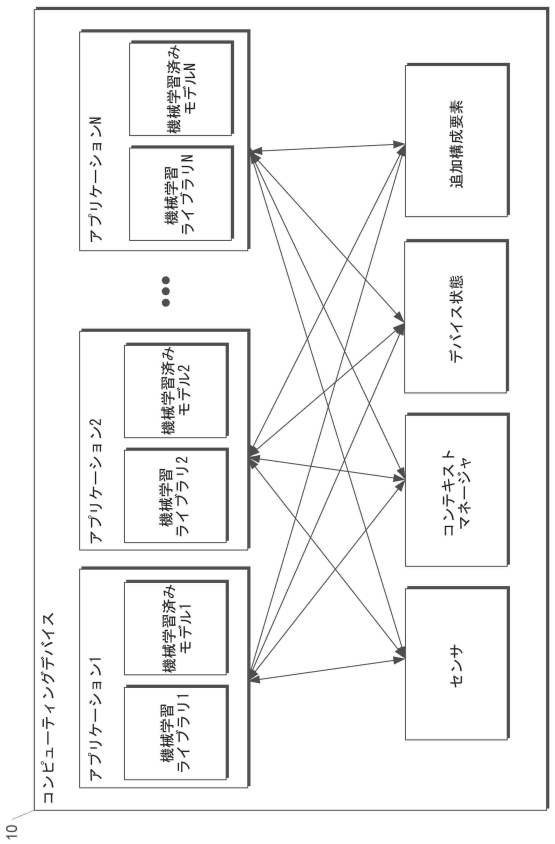
【図４】


418
416
出力トークン
報酬関数
414
構造ツール
機械学習済み
言語モデル
415
文脈トークン
412

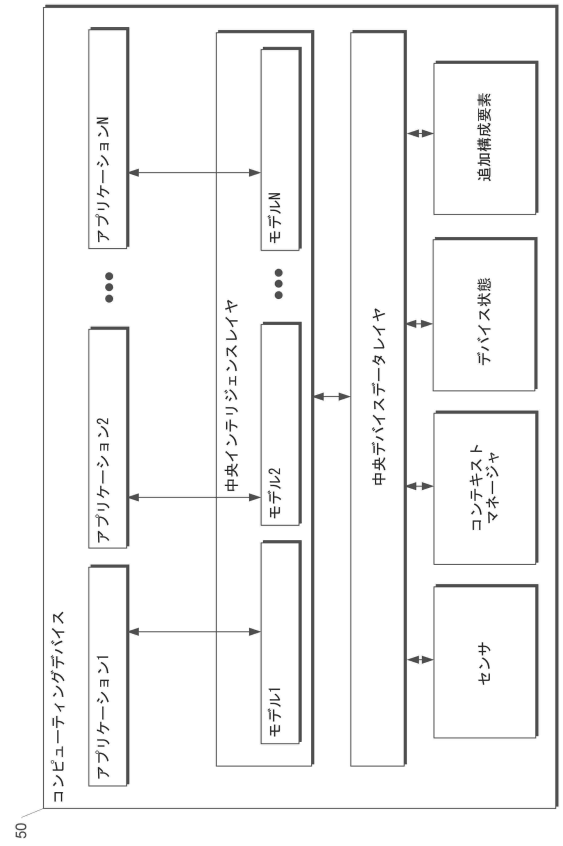
【図５Ａ】


サーバコンピューティングシステム
プロセッサ
メモリ
データ
命令
機械学習済みモデル
トレーニング用コンピューティングシステム
プロセッサ
メモリ
データ
命令
モデル訓練器
訓練データ
ユーザコンピューティングデバイス
プロセッサ
メモリ
データ
命令
機械学習済みモデル
ユーザ入力構成要素
100
102
130
132
134
136
138
140
150
152
154
156
158
160
162
180

【図５Ｂ】


コンピューティングデバイス
アプリケーション1
機械学習ライブラリ1
機械学習済みモデル1
アプリケーション2
機械学習ライブラリ2
機械学習済みモデル2
アプリケーションN
機械学習ライブラリN
機械学習済みモデルN
センサ
コンテキストマネージャ
デバイス状態
追加構成要素
10

【図５Ｃ】


コンピューティングデバイス
アプリケーション1
アプリケーション2
アプリケーションN
中央インテリジェンスレイヤ
モデル1
モデル2
モデルN
中央デバイスデータレイヤ
センサ
コンテキストマネージャ
デバイス状態
追加構成要素
50

フロントページの続き

(72)発明者　ノーム・シェイザー
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００・グーグル・エルエルシー内
(72)発明者　ダニエル・デ・フレイタス・アディワルダナ
アメリカ合衆国・カリフォルニア・９４０４３・マウンテン・ビュー・アンフィシアター・パークウェイ・１６００・グーグル・エルエルシー内

審査官　齊藤 貴孝
(56)参考文献　特表２０２１－５２４６２３（ＪＰ，Ａ）
米国特許出願公開第２０１９／０１３０３０５（ＵＳ，Ａ１）
米国特許出願公開第２０２０／０１１０８０３（ＵＳ，Ａ１）
米国特許出願公開第２０２１／００７３４６５（ＵＳ，Ａ１）
栗林 樹生、外６名，論述構造解析におけるスパン分散表現，自然言語処理，日本，一般社団法人言語処理学会，2020年12月15日，第２７巻，第４号，p.７５３－７７９
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　４０／００－４０／５８