AI 特許紹介(51)
AI 特許を学ぶ！究める！
～コントローラブル GPT 特許～

2023 年 4 月 10 日

河野特許事務所
所長弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要

特許出願人　Microsoft
出願日　2020 年 3 月 12 日
公開日　2021 年 6 月 24 日
公開番号 US20210192140
発明の名称　コントローラブルグラウンディングテキスト生成

140 特許は、GPT(Generative Pre-trained Transformer)等の言語生成モデルにグラウンディングソースとコントロールシグナルを与えることにより、コントローラブルなグラウンディング応答生成（CGRG：Controllable Grounded Response Generation）を実現する技術に関する。

2.特許内容の説明

人間の話者と区別がつかないテキストを生成することは非常に難しい問題である。GPT-2 モデルなどの大規模ニューラル生成モデルに関する最近の研究では、人間の話者に由来するテキストにより近いスタイルとフローを持つテキストを生成する可能性

が示されている。しかしながら、このようなニューラルモデルによって生成されたテキストを詳しく調べてみると、無意味なステートメントや文脈上誤った事実が含まれていることがよくある。

下記図 1 は、機械学習モデル 102 を実装するテキスト生成コンピューティングシステム 100 のブロック図である。

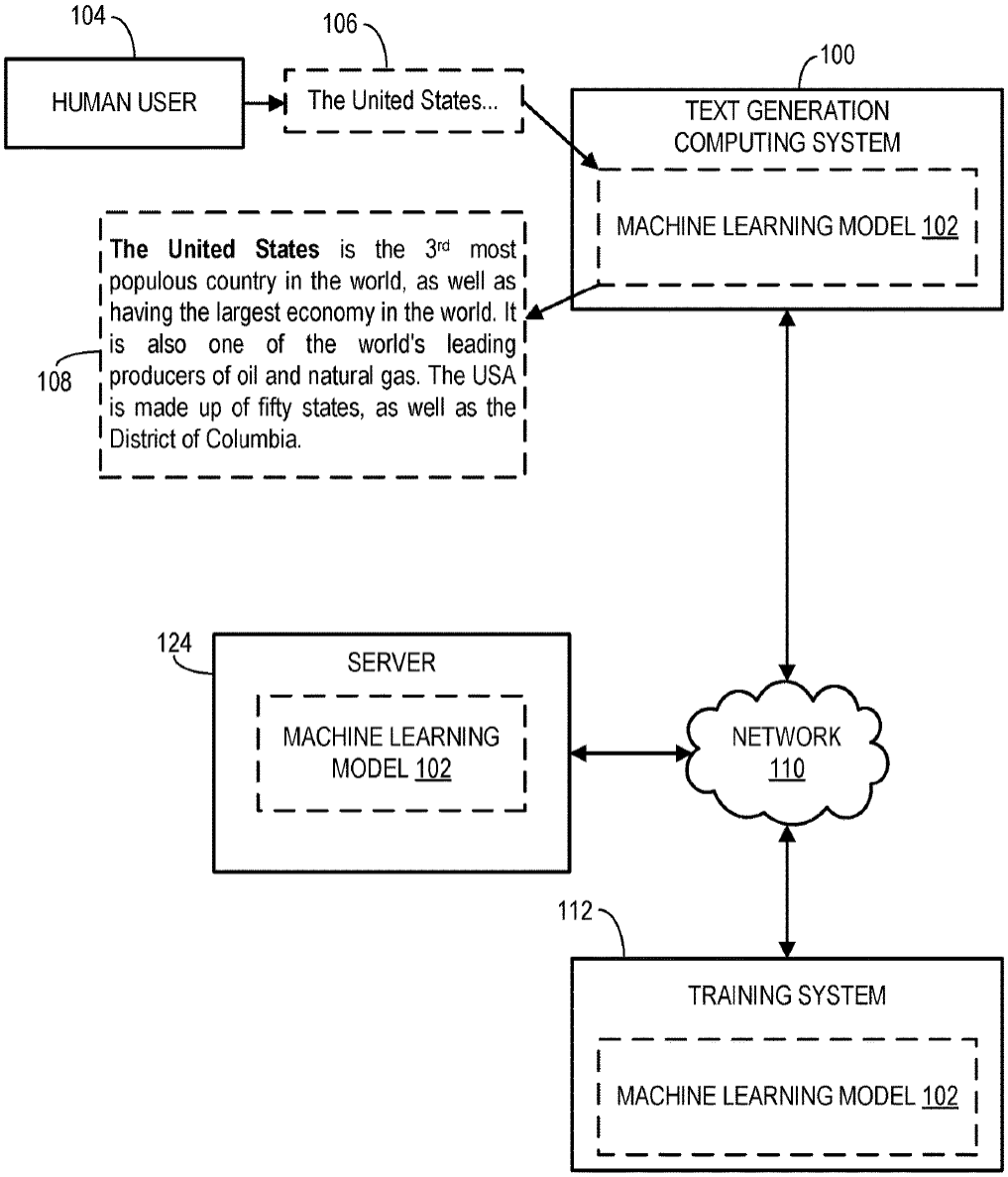


図 1 の例において、ユーザ 104 がテキスト 106「The United States」を入力した後、テキスト生成コンピューティングシステムは、機械学習モデル 102 を使用して、コンピュータ生成テキスト 108 を出力する。コンピュータ生成テキスト 108 は、ユーザ提供テキスト 106 の「The United States」トピックを拡張し、ユーザ提供テキスト 106 に

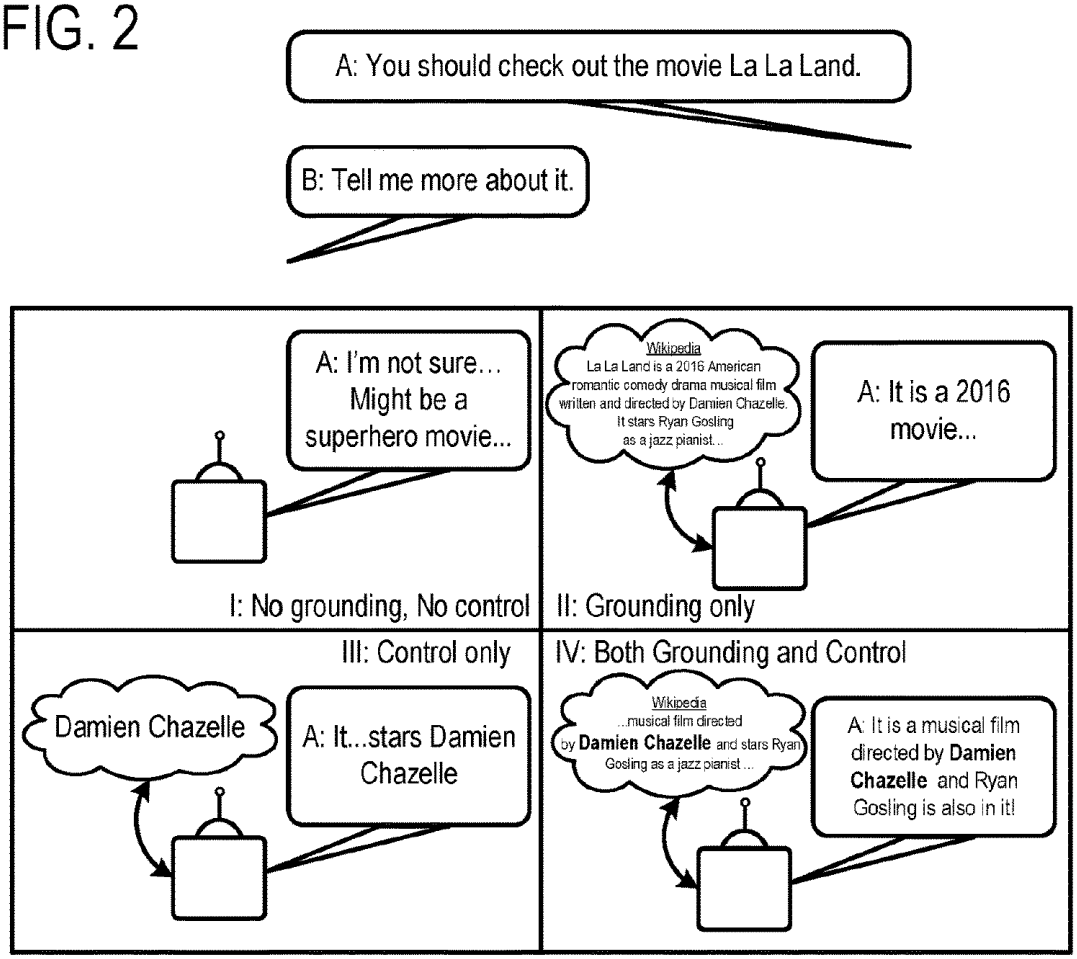
関連する短い文章を自動的に生成する。

ネットワーク 110 には、機械学習モデル 102 を訓練するように構成された訓練システム 112 も結合される。機械学習モデル 102 は、トレーニングシステム 112 でトレーニングされ、その後、テキスト生成コンピューティングシステム 100 またはサーバ 124 に展開される。機械学習モデル 102 は、GPT-2 または GPT-3 等のトランスフォーマベースの言語モデルを含み、セルフアテンションを使用する。

オープンドメイン応答生成用のエンドツーエンドニューラルモデルは、流暢で文脈的に適切な会話応答を生成できる。初期のニューラル生成モデルは当たり障りのない回避的な反応が特徴であったが、驚くべきことに、最近の多様性を強化する戦略と大規模な GPT-2/GPT-3 スタイルモデルを使用して、人間のような会話を生成できる。

下記図 2 はシナリオ例を示す説明図である。マイナス面として、図 2 のシナリオ I に示されている種類の「幻覚」または「偽の」出力への傾向が存在する。例えば、ユーザが「それについてもっと教えてください」と述べ、ここでモデルが「よくわかりません。. . スーパーヒーロー映画かもしれません。 . . 」と反応する。

FIG. 2



グラウンディング（根拠のある）反応生成アプローチは、事実の幻覚を抑制することができる。しかし、コントロールまたはセマンティックターゲティングを行わずにグラウンディングだけを行うと(例えば、図 2 のシナリオ II の「ラ・ラ・ランド」に関するウィキペディアのページ)、正確ではあるが曖昧または無関係なアウトプットを誘発する可能性がある。例えば、図 2 のシナリオ II のように、モデルが「それは 2016 年の映画です・・・」と反応してしまう。

一方、コントローラブルなテキスト生成は、デコーダを関連する出力に導くことができるレベルのセマンティック制御を提供するが、根拠となる（グラウンディング）制御フレーズがなければ、正しい事実に関連付けることができない。（例えば図 2 のシナリオ III では、モデルは次のように応答する。「主演は Damien Chazelle ダミアン・チャゼル」）

本発明のテキスト生成フレームワークは、グラウンディング知識と語彙コントロールの両方を組み込んで、信頼できる文脈に適した情報を備えた人間に似たテキストを生成

する。例えば図 2 のシナリオ IV では、モデルは次のように応答する「Damien Chazelle 監督のミュージカル映画で Ryan Gosling も出演！」

図 3 は処理ロジックを示す説明図である。

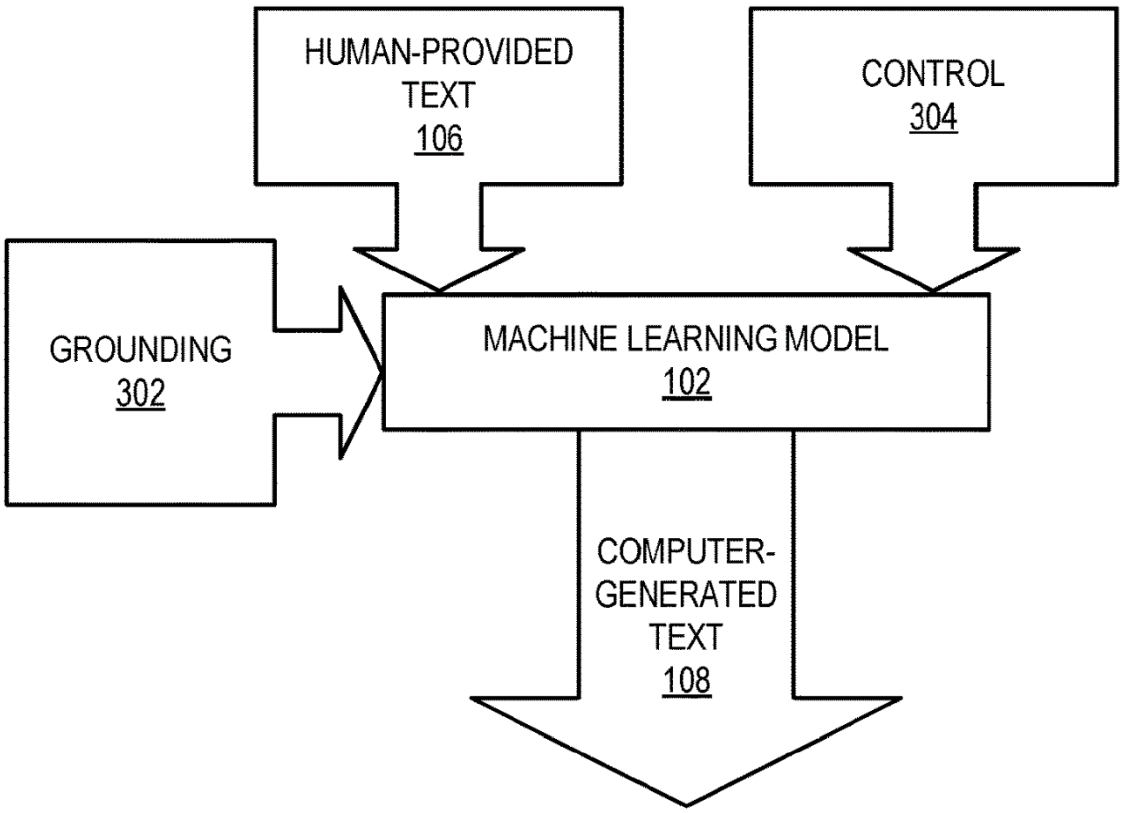


図 3 に示されるように、制御可能なグラウンディング応答生成フレームワーク 300 は、機械学習モデル 102 を使用して、1）人間が提供するテキスト 106、2）グラウンディング 302、および 3）コントロール 304 に基づいてコンピュータ生成テキスト 108 を出力する。

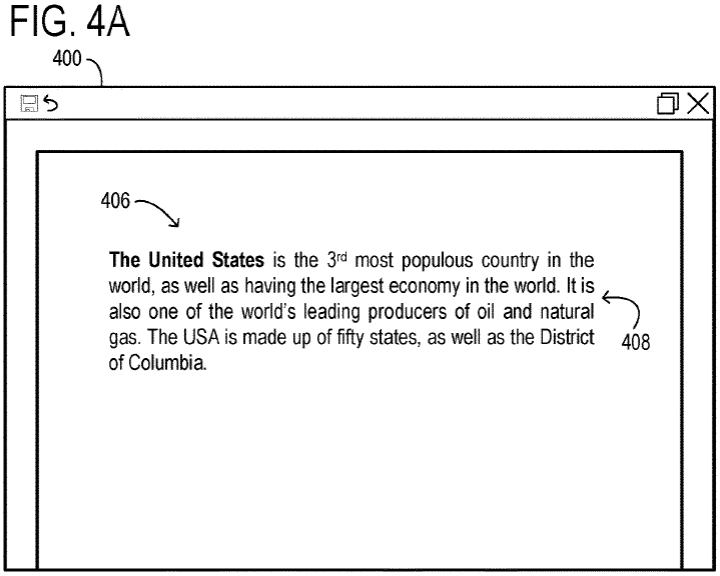
これらのインターフェースを使用して、機械学習モデルは、グラウンディングソースから情報を取得し、コントロールシグナルに基づいてコンピュータで生成されたテキストに焦点を合わせる。グラウンディング 302 とコントロール 304 の両方を使用することにより、コンピュータ生成テキスト 108 は、グラウンディングまたはコントロールのみが利用された場合に生成されるよりも、より高い品質（たとえば、より高いコンテキストの関連性、より高い事実上の精度、ユーザの利益により焦点を合わせている）となる。

グラウンディング 302 には、マシンアクセス可能なデータベースおよびその他の情報ストアで収集されたドメイン存在およびドメイン固有の情報が含まれる。例えば、グラウンディング 302 は一般的または特定の検索エンジンを利用する。

コントロール 304 には、コンテンツプランナーまたはその他の自動化されたシステムから人間が提供するコントロールまたは自動的に抽出されたコントロールが含まれる。たとえば、人間のユーザがワードプロセッサを使用してドキュメントを作成しているシナリオでは、ワードプロセッサをユーザから入力を受信し、受信した入力をコンピュータで生成したテキストに焦点を合わせるためのコントロール信号として使用するように構成できる。ユーザが提供するコントロールは、その人が最も興味深く、または適切であると考える根拠となる事実に文章の内容を集中させることができる。

図 4 はコントロールシグナルの例を示す説明図である。

FIG. 4A



例えば、図 4A は、ユーザが、テキスト 406「The United States」をワードプロセッサ 400 にタイプし、ワードプロセッサが、１つまたは複数のグラウンディングソース(たとえば、図 3 のグラウンディング 302-たとえば、米国に関するウィキペディアの記事)を活用する機械学習モデルを使用して、コンピュータ生成テキスト 408 を表示するシナリオを示している。

FIG. 4B

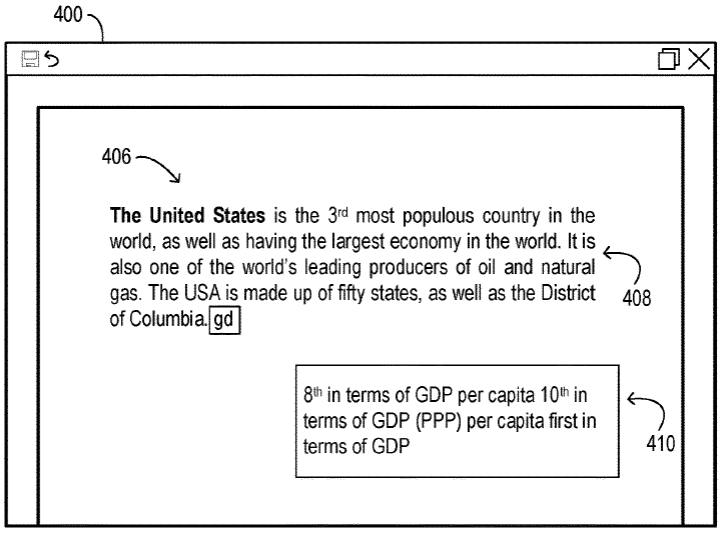


FIG. 4C

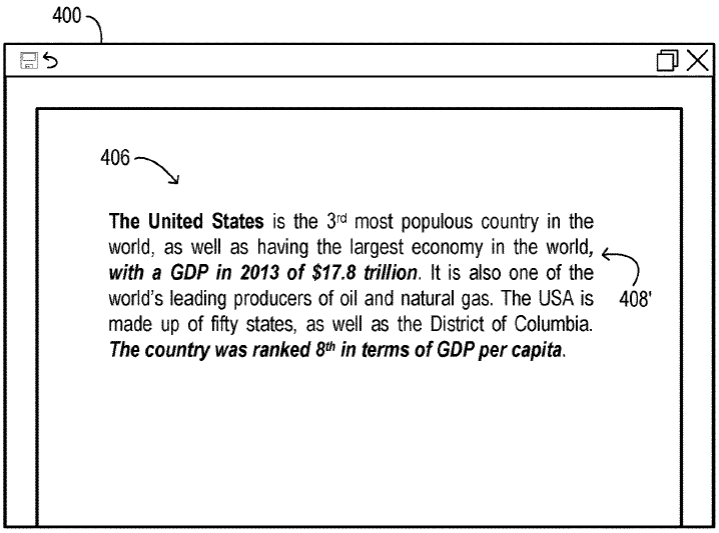


図 4B は、図 4A の例の続きであり、コンピュータ生成テキスト 408 の続きとして、ユーザが文字「gd」をタイプしたことを示している。「gd」の認識に応答して、ワードプロセッサは、コントロールシグナルによって制御されるように、「The United States」グラウンディングソースから得られた顕著な事実 410 を示す。

図の例では、ワードプロセッサは 3 つの顕著な事実を提示し、ユーザは最初に提示された事実、つまり「1 人あたりの GDP で 8 位」を選択する。選択された「一人当たり GDP で 8 位」は、機械学習モデルへのコントロールシグナルとして提供される。

図 4C は、更新されたコンピュータ生成テキスト 408 を示しており、GDP コントロールシグナルに基づいて追加された文章は太字および斜体で示されている。

FIG. 4D

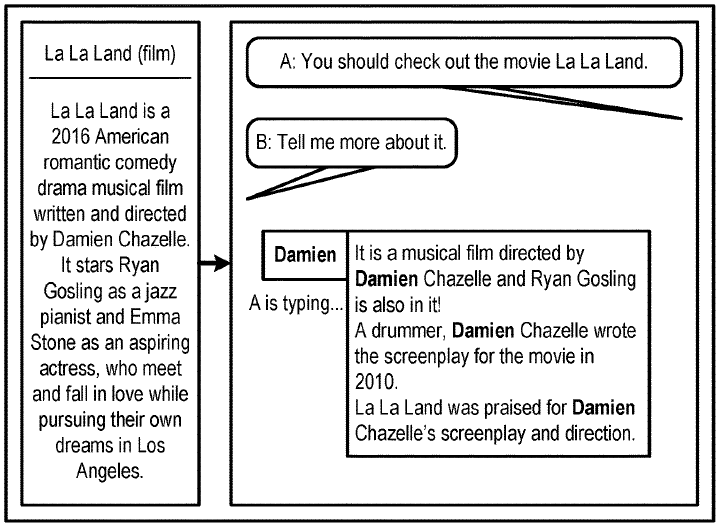
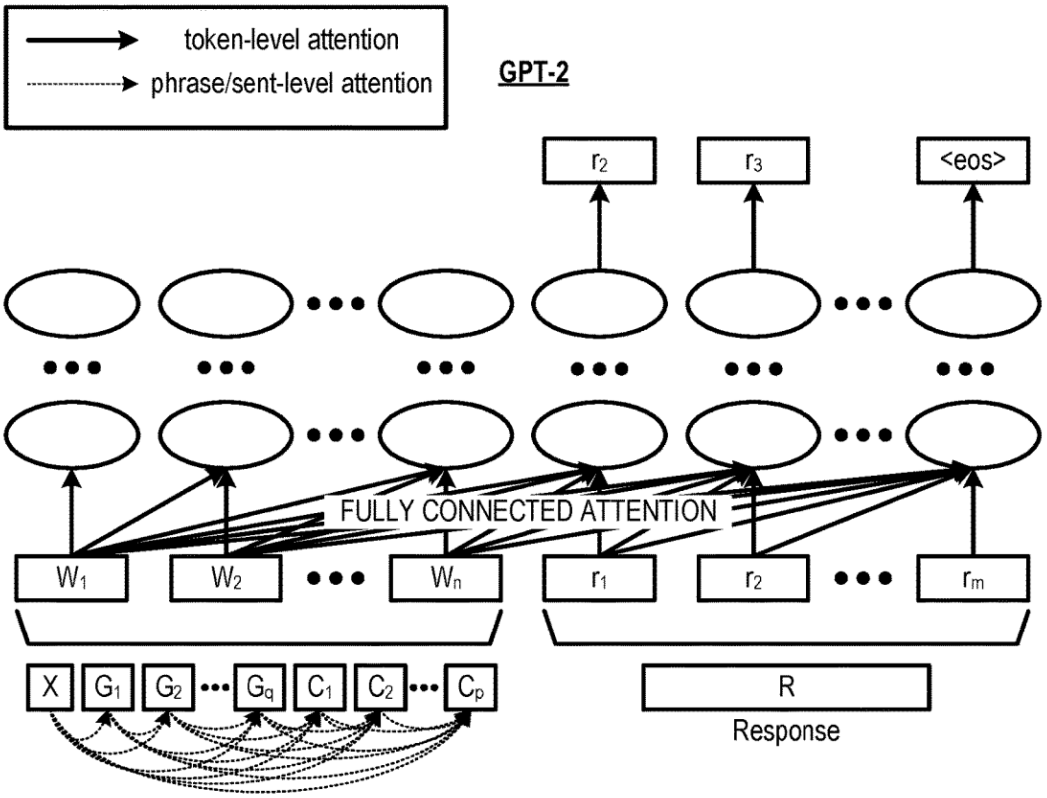


図 4D は、人が意味的意図を示すためにキーワードを入力し、機械学習モデルがコントロールを使用してコンピュータ生成テキストを出力する別の例を示している。特に、機械学習モデルは、会話履歴、ユーザの部分的な入力（Damien)、およびグラウンディングとなる知識に従って、ユーザ A の候補応答を提案する応答編集アシスタントとして機能する。

図 5 は機械学習モデルのネットワーク構成を示す説明図である。



制御可能なグラウンディング応答生成フレームワークの概念は、次のように形式化できる。与えられた対話コンテキスト X、複数の(p)語彙コントロールフレーズ $C=(C_1, \ldots, C_p)$およびグラウンディング $G=(G_1, \ldots, G_q)$の q 文は、C によって導かれるセマンティック情報が含まれる応答 $R=(r_1, \ldots, r_m)$を生成する。コントロールは、ユーザが直接提供することも、コンテンツプランナーから自動的に取得することもできる。区別するために、検証済みまたはユーザ提供のコントロールは C として示され、コンテンツプランナーによって提供されるコントロールは $\tilde{C}$として示される。

制御可能なグラウンディング応答生成は、必要に応じて、グラウンディング会話データセットと連携して使用できる。人間がすべてのコントロールフレーズに注釈を付けることはコストがかかり、かつ拡張できない可能性があるため、語彙マッチングが使用される。

機械学習モデルは、GPT-2 機械学習モデルを含むか、または GPT-2 機械学習モデルから導出することができる。GPT-2 は、大規模な Web データでトレーニングされたトランスフォーマベースの言語モデルであり、各トークンが残りのトークンにアテンションするセルフアテンションを使用する。これは、定義されたコンテキストウィンドウ内

の前のすべての単語を考慮して、次の単語を予測することを目的としてトレーニングされる。

CGRG 内で GPT-2 を適用するために、図５上に示されるように、X,C（および／またはC⁻）およびG_Cが入力シーケンスとして連結される。このモデルは、連結された入力シーケンス（S として示される）と R 内の前の応答トークンが与えられると、次の応答単語を予測する。G_Cは、C に関連する G のサブセットである。

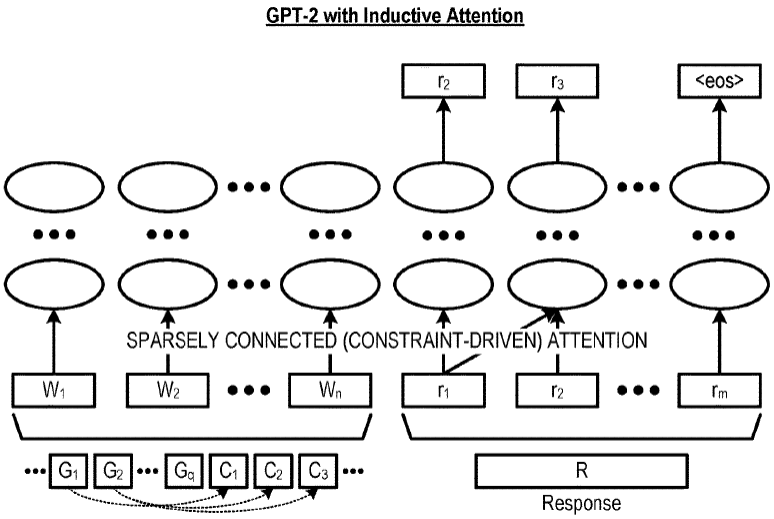
たとえば、本例では、Care の任意のフレーズを含むグラウンディングセンテンスをG_Cと表記している。入力要素を区別するために、テキスト終了トークン <eos> が X の各対話発話の終わりに挿入され、<c> トークンが C の各コントロールフレーズの最後に挿入され、<s> トークンが G_Cの各文の最後に挿入される。

入力シーケンス S と応答シーケンス R は、最初に長いテキストに連結される。
ソースシーケンスは$S=(w_1, \ldots, w_n)$として表され、ターゲットセンテンス R を生成するために使用される。$P(R|S)$の条件付き確率は、条件付き確率の積として記述できる。

$$p(R \mid S) = \prod_{k=1}^{m+1} p(r_k \mid w_1, \ldots, w_n, r_1, \ldots, r_{k-1})$$

ここで、r_{m+1}は、生成の終わりを示す追加のテキスト終了トークンである。

デフォルトでは、GPT-2 は連続したテキストシーケンスを入力として受け取る。上記のアプローチを使用すると、X、C、GC の各入力要素はセグメント化された形式であり、これらのセグメントは必ずしも強く接続されているとは限らない。したがって、単純にすべてを GPT-2 モデルに連結すると、ノイズが発生する可能性がある。



C と G_C の間に事前に確立された構造情報を注入することにより、各データ例の潜在的に有益でないアテンションリンクを削除できる。例えば、図５下では、C は C_1, C_2, C_3 を含み、G_C は G_1 および G_2 を含む。

C_1 が G_1 のみにあることがわかっている場合は、C_1 と G_1 の間のアテンションリンクのみを保持する必要があり、C_1 と他のグラウンディング文章との間では保持しないようにする。G_C は G から分割された文の集合であると考えられているため、G_C トークン内の文間のリンクはすべて削除されている。

同様に、同一でないフレーズ間のすべてのリンクが削除される。したがって、各データ例のアテンションリンクは、C と G_C の間の構造情報によって事前に決定される。これを実装するために、各トランスフォーマレイヤで、削除されたアテンションリンクと将来のトークンへのリンクの値が 0 で、その他のリンクの値が 1 であるアテンションマスクが適用される。この事前に計算されたアテンションは、誘導的アテンションと呼ばれる。各応答トークンは、左側にあるすべての入力トークンとその他の応答トークンに引き続き対応する。

S 中のコントロール句 $C_i \in C$ の開始位置と終了位置を c_i^s と c_i^e で表し、グラウンディングセンテンス $G_i \in G_C$ の開始位置と終了位置を g_i^s と g_i^s で表す。次に、アテンションマスク M は次のように計算される。

$$M_{i,j} = \begin{cases} 0 & \text{if } i < j \\ 0 & \text{if } i \in [c_k^s, c_k^e],\ j \in [c_l^s, c_l^e],\ k \neq l \\ 0 & \text{if } i \in [g_k^s, g_k^e],\ j \in [g_l^s, g_l^e],\ k \neq l \\ 0 & \text{if } i \in [c_k^s, c_k^e],\ j \in [g_l^s, g_l^e],\ C_k \notin G_l \\ 1 & \text{otherwise} \end{cases}$$

次に、各トランスフォーマヘッドについて、積み重ねられた行列 Q、K および V は、各シーケンス例（連結された S および T）を表すことができる。アテンションは次のように計算される（dis はモデルの次元である）。

$$\text{Attention}(Q, K, V) = \text{softmax}\frac{M \circ QK^T}{\sqrt{d}}V$$

3.クレーム

140 特許のクレーム 1 は以下の通りである。

1.制御可能なグラウンド応答生成フレームワークをインスタンス化するためにロジック サブシステムによって実行可能な命令を保持するストレージサブシステムにおいて、

入力テキストに基づいてコンピュータ生成テキストを出力するようにトレーニングされた機械学習モデルと、

入力テキストに関連する情報を含むグラウンディングソースにアクセスするために機械学習モデルで使用可能なグラウンディングインターフェースと、

コントロールシグナルを認識するために機械学習モデルによって使用可能な制御インターフェースと、

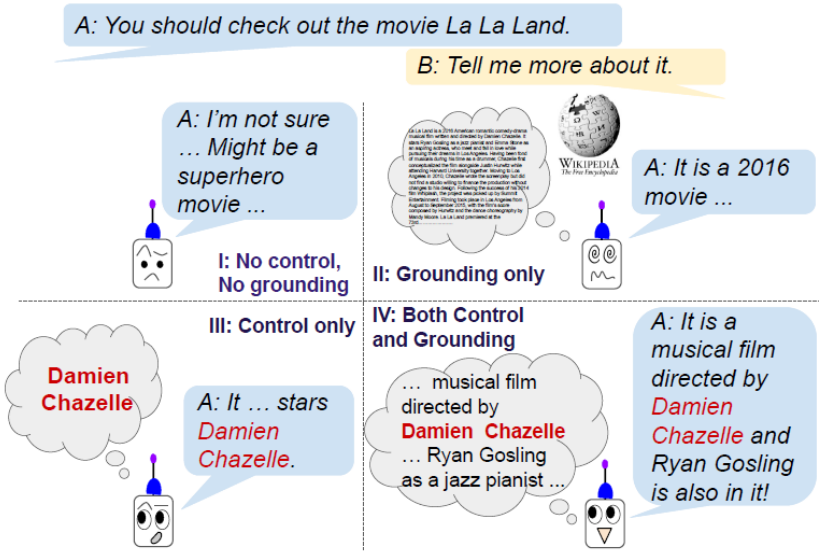
機械学習モデルは、グラウンディングソースからの情報をコンピュータ生成テキストに含め、コントロールシグナルに基づいてコンピュータ生成テキストに焦点を当てるように構成されている。

4. 本特許に関連する論文

本特許に関する論文“A Controllable Model of Grounded Response Generation”[1]

[1] ZeqiuWu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, Bill Dolan “A Controllable Model of Grounded Response Generation” arXiv:2005.00613v2 [cs.CL] 14 Jun 2021

が、Microsoft の ZeqiuWu 氏らにより公表されている。



140 特許で説明されていた文例が示されている。生成された応答は、一般的なものであったり、グラウンディングまたはコントロールがなければ事実に反するものである傾向がある（I)。グラウンディングを追加すると、情報の信頼性は向上するが、あいまいな応答につながる可能性がある(II)。コントロールを追加すると応答の特異性は高まり(III)、両方を使用すると、満足のいく信頼できる応答が得られる(IV)。

下記表は対比結果を示す説明図である。

Setting	Model	NIST	BLEU	Div-2	Avg-L
1) X	GPT-2	0.90	0.55%	4.9%	22.2
2) X+G	CMR	0.34	0.17%	11.3%	15.1
3) X+G	GPT-2	0.98	0.67%	7.5%	23.1
4) X+C	GPT-2	1.67	2.65%	10.7%	28.7
5) X+G_C	GPT-2	1.34	1.58%	11.1%	26.6
6) X+C+G	GPT-2+GBS[7]	1.60	2.38%	10.6%	26.8
7) X+C+G_C	GPT-2	1.77	3.22%	11.3%	27.0
8) X+C+G_C	GPT2IA	**1.80**	**3.26%**	**11.6%**	25.9

上記表において、行１～３はコントロール可能な設定ではなく、入力としてコントロール句を持たないが、行４～８は明示的または暗示的に入力としてコントロール句を有する。

行（1-3）と（4-8）の間の大きなパフォーマンスギャップは、コントロールを追加することの価値を示している。さらに、相互の比較により、以下の結論を導き出すことが

できる。
(i) 1 vs. 3: モデル入力に単にグラウンディングを追加するだけで、限られた範囲でパフォーマンスが向上する
(ii) 2 vs. 3: 一般に、GPT-2 は最先端のグラウンディングモデル CMR(Continual Model Refinement)よりも優れたパフォーマンスを発揮する。これは、事前トレーニングとトランスフォーマベースのデコーダーの組み合わせがテキスト生成の改善に役立つことを示している。
(iii) 4 vs. 7-8: 制約に敏感なグラウンディングを提供すると、すべてのグラウンディングを行う場合と比較してパフォーマンスが向上する
(iv) 5 vs. 7-8: 明示的な方法でコントロールフレーズを提供することが重要である。
(v) 6 vs. 7-8: 隠れ状態にコントロールを適用すると、デコードのみにコントロールを適用するよりも、モデルがより質の高い応答を生成するのに役立つ
(vi) 7 vs. 8: 誘導性アテンション（IA : Inductive attention）は、ノイズを低減し、GPT-2 のパフォーマンスを向上させるのに役立つ。

以上

著者紹介
河野英仁
河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。
AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 3」、「ブロックチェーン 3.0(共著)」がある。