

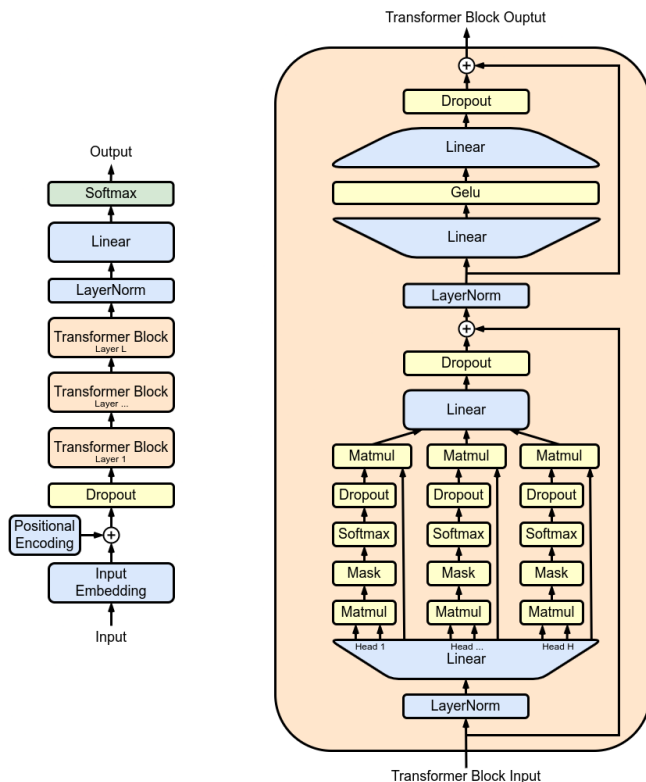
大規模言語モデルと著作権に関する 一考察

情報通信研究機構

鳥澤 健太郎

2023年10月16日


- Transformerというアーキテクチャに基づく、超巨大なニューラルネット
- 入力されたテキストの次の単語を確率的に予測するだけ→言語「モデル」と呼ばれる所以
- 以下の説明はGPTと呼ばれる大規模言語モデルに関してのみ。今後、別種の大規模言語モデルが出現した場合には再度検討が必要になる可能性

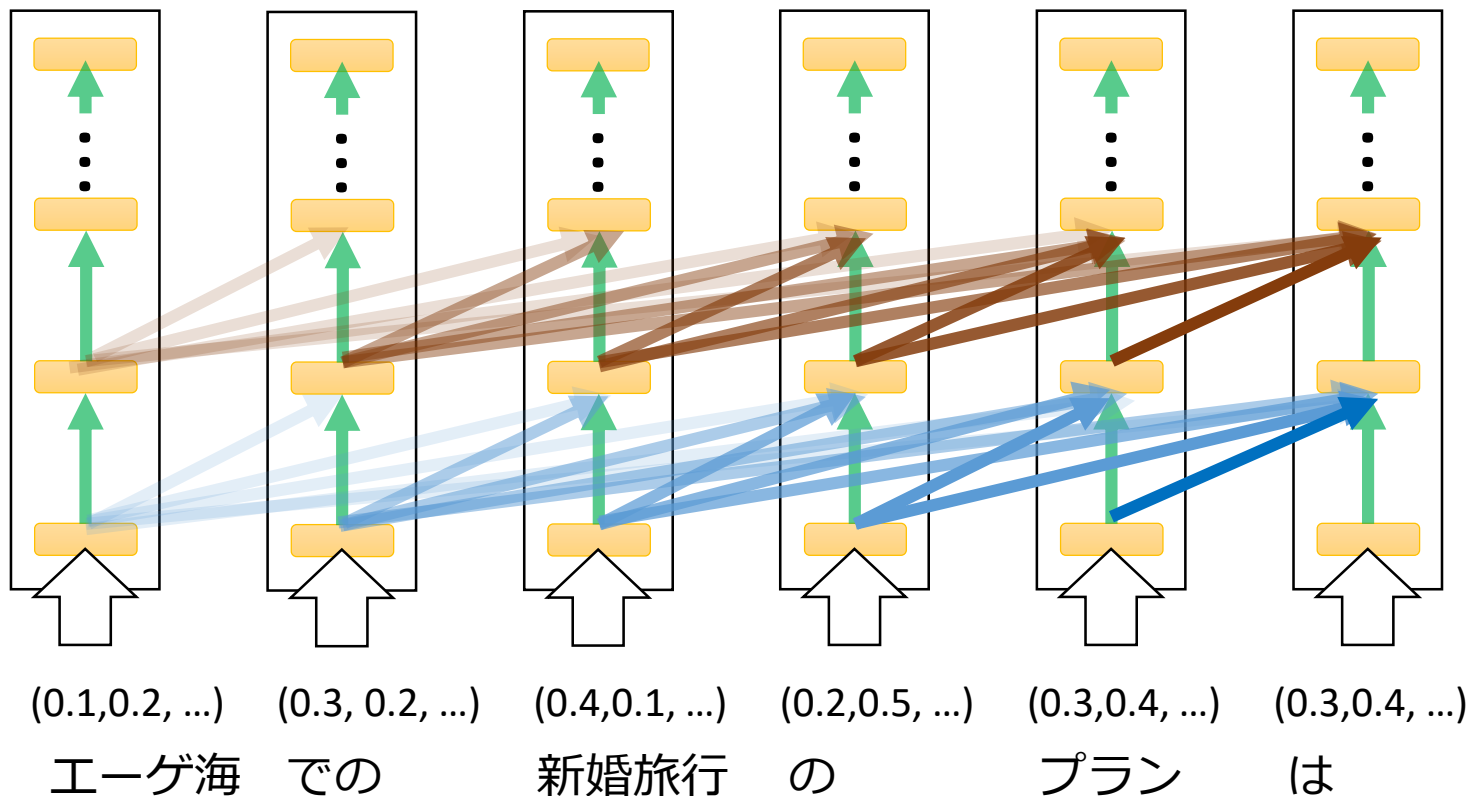


Wikipedia日本語版の
「GPT(言語モデル)」の項より転載
[https://ja.wikipedia.org/wiki/GPT_\(言語モデル\)](https://ja.wikipedia.org/wiki/GPT_(言語モデル))

CC-BY-SA 4.0

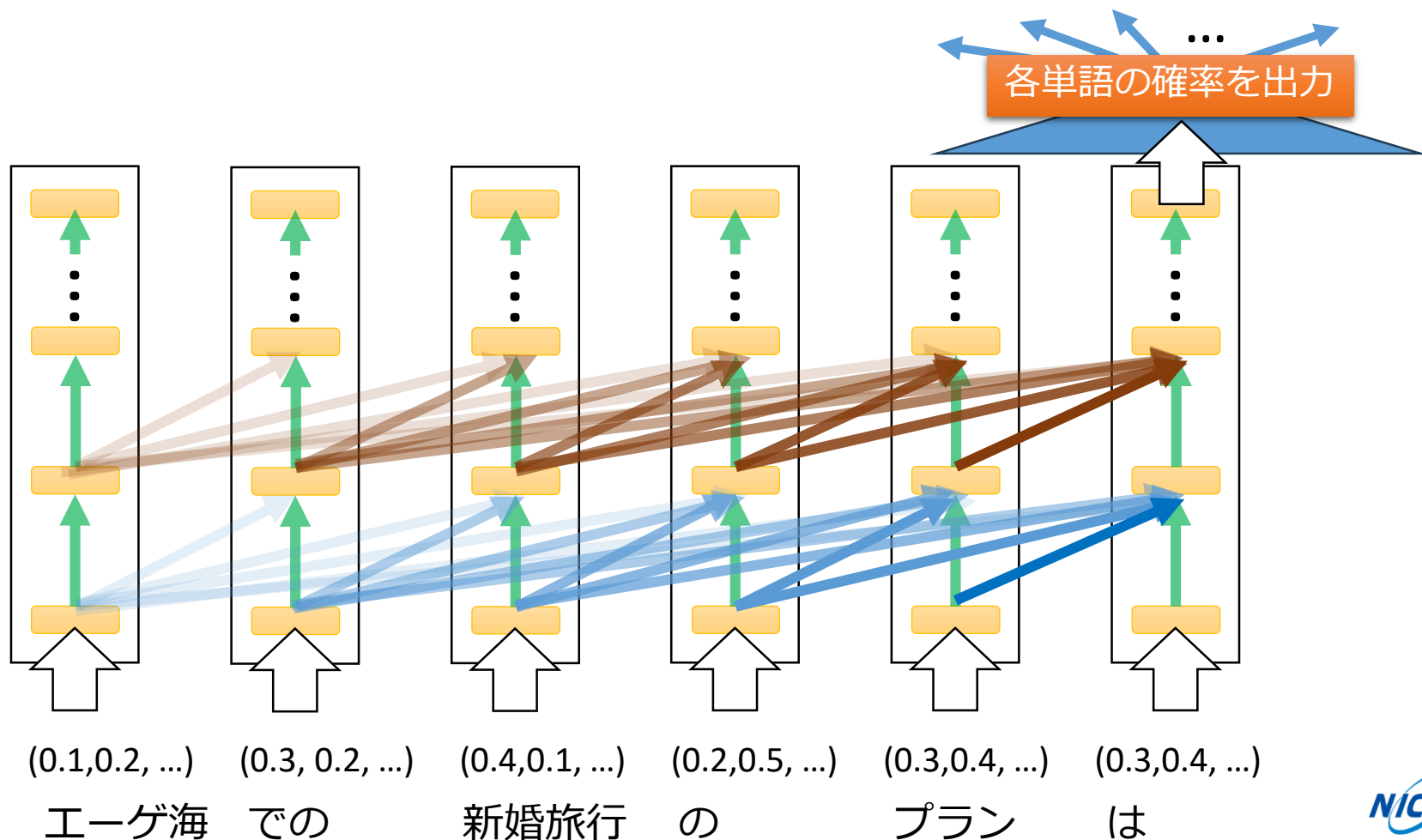
<https://creativecommons.org/licenses/by-sa/4.0/deed.ja> に則り転載

- 「エーゲ海での新婚旅行のプランは」と入力してみると、その続きのテキストを生成
- 単語はベクトル (= ) に変換
- ニューラルネットの下から上に向けて、単語のベクトルを変形しつつ、周囲の単語のベクトルを融合することを（数十回から数百回）繰り返す



• 続きの単語を生成するには？

- 一つ前の単語（この場合は「は」）に関する出力のベクトルをもとに、辞書にあるすべての単語が次の単語として現れる確率を計算
- 基本的には確率最大の単語を「次の単語」として生成

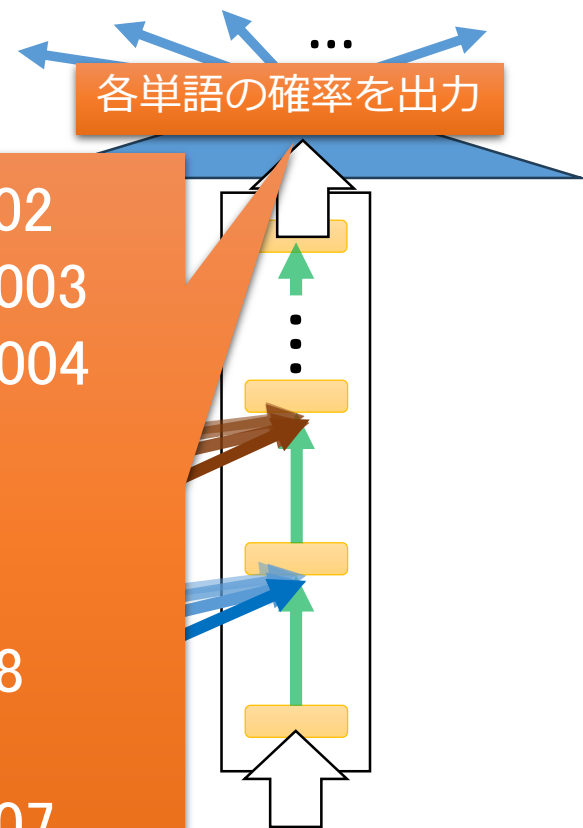


• 続きの単語を生成するには？

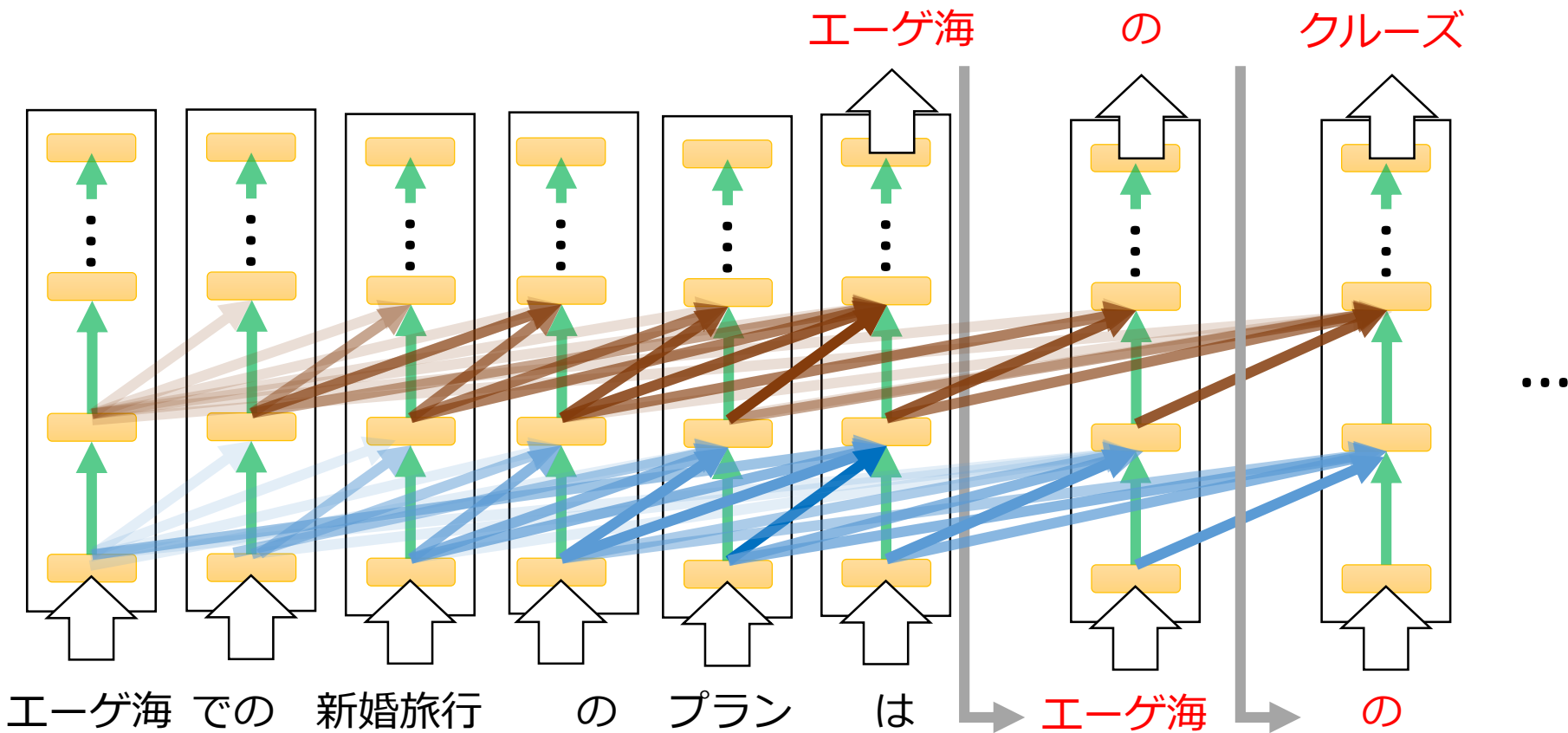
- 一つ前の単語（この場合は「は」）に関する出力のベクトルをもとに、辞書にあるすべての単語が次の単語として現れる確率を計算
- 基本的には確率最大の単語を「次の単語」として生成

辞書にある数万語から数十万語のすべての単語について確率を計算。基本的には最大のもの（この例では「エーゲ海」）を選ぶ

「あ」の確率 =	0.00002
「あい」の確率 =	0.000003
「愛」の確率 =	0.000004
...	
「エーゲ海」の確率 =	0.001
...	
「クルーズ」の確率 =	0.0008
...	
「、」の確率 =	0.00007
テキスト終了の確率 =	0.0006

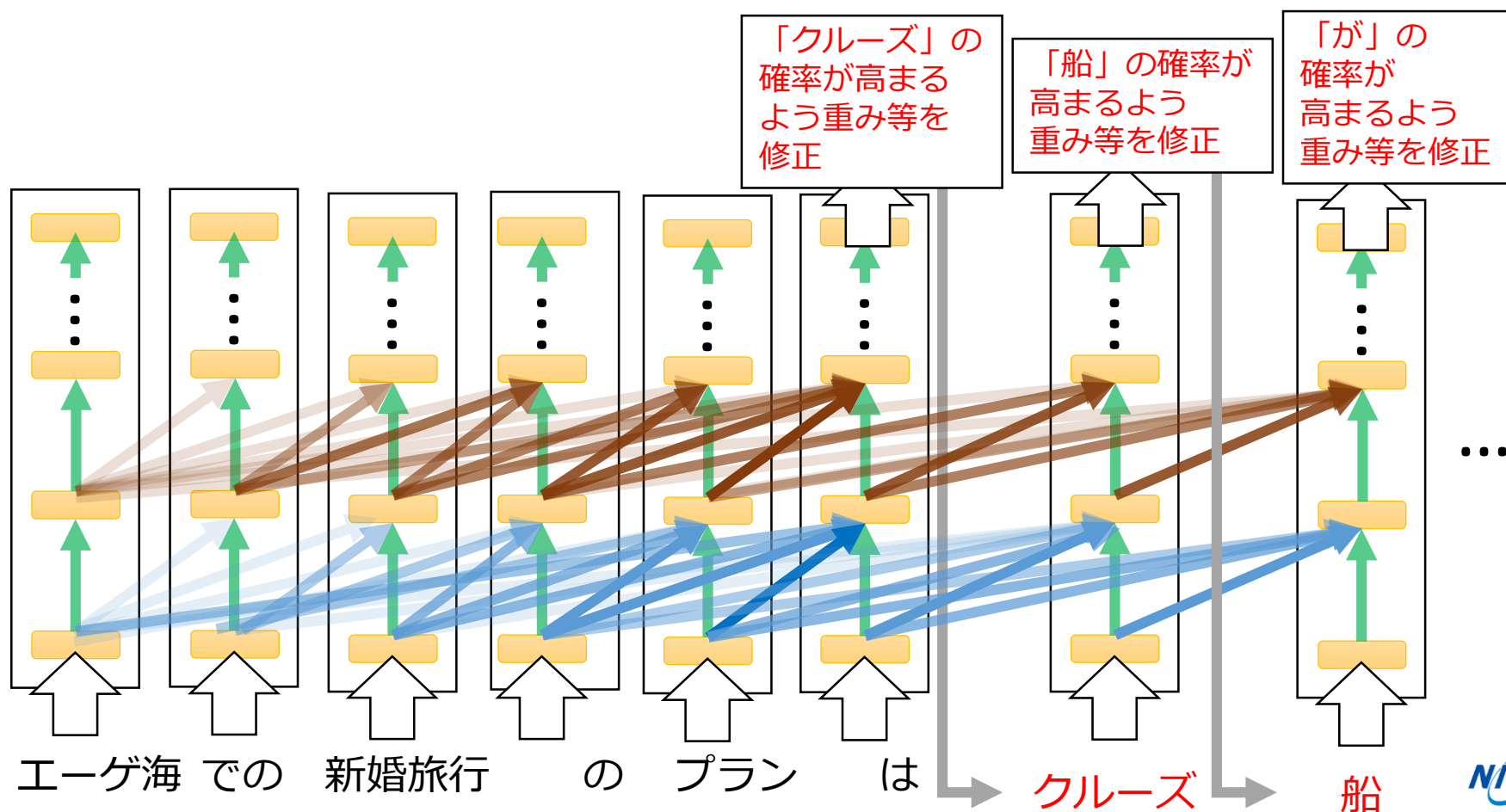


- 基本的には、以上の動作を繰り返すことで長いテキストを生成
 - 生成された単語は入力の末尾に追加される



学習はどのようにするのか？

- 学習データの「正解テキスト」中の単語の確率が高くなるように、ニューラルネットの重みや単語ベクトルを修正
- これを大量の正解テキストに対して繰り返す
- 例えば、学習データ中の正解テキストが「エーゲ海での新婚旅行のプランはクルーズ船がおすすめ」であったとすると、



学習はどのようにするのか？

- 学習データの「正解テキスト」中の単語の確率が高くなるように、ニューラルネットの重みや単語ベクトルを修正
- これを大量の正解テキストに対して繰り返す
- 例えば、学習データ中の正解テキストが「エーゲ海での新婚旅行のプランはクルーズ船がおすすめ」であったとすると、

「クルーズ」の
確率が高まる
よう重み等を
修正

「船」の確率が
高まるよう
重み等を修正

「が」の
確率が
高まるよう
重み等を修正

- 大量の正解テキストを参照しながら単語の確率を徐々に大きくしていく
- 特定の一つの正解テキストのみに基づいてある特定単語の確率を直ちに最大にするわけではない（つまり、正解テキストをそのまま出力することを学習させているのではない）
- このため、仮に学習データ中のテキストXがあたかも丸ごとコピーされて出力されたように見えても、Xを用いた学習のプロセスとXが丸ごと出力されることの間因果関係は不明確
 - 逆にいうと、誰かが書いた、学習データ中にないテキストXが、Xとは異なる他のテキストから学習した結果、出力されることもあり得る。「誰かが書いた」という条件を除けば、大規模言語モデルのユーザであれば誰でも日常経験していることではあるが。

学習はどのようにするのか？

- 学習データの「正解テキスト」中の単語の確率が高くなるように、ニューラルネットの重みや単語ベクトルを修正
- これを大量の正解テキストに対して繰り返す
- 例えば、学習データ中の正解テキストが「エーゲ海での新婚旅行のプランはクルーズ船がおすすめ」であったとすると、

「クルーズ」の
確率が高まる
よう重み等を
修正

「船」の確率が
高まるよう
重み等を修正

「が」の
確率が
高まるよう
重み等を修正

- 学習が終了した時、単語のベクトルはその単語の用法や意味を反映していると考えられるが、具体的にどのように反映しているかを知るのは極めて困難
- ネットワーク内の処理、学習はベクトル、数値計算だけで行われ、いわゆる文字列の切り貼り等の操作は一切なし
- AIの研究を30年やってきた者からしても、むしろ意味の通るテキストがこのような学習で出力される方がむしろ不思議
- 常に、というわけではないが、抽象的な概念、アイデアまで学習し、それに基づいてテキストを生成しているという「印象」

エーゲ海 での 新婚旅行 の プラン は クルーズ 船

一見、学習データがそのまま出力されるように
思われるかもしれないが、

- 実際には非常に大量の学習データで学習
- 一旦生成できるようになったテキストを学習が進むと忘れてしまうことも

学習ステップ	学習データ
0	卒業旅行でエーゲ海に行きました。自分でフェリーを...
...	...
12345	エーゲ海での新婚旅行のプランはクルーズ船がおすすめ。
...	...
25687	クルーズ船で感染症が発生すると、高齢者が多数感染して、
...	...
67452	アメリカ人ですが、奈良で新婚旅行をしてきました。鹿が.....

仮に、この段階で 学習データのテキストを丸ごとコピーして出力するようになったとしても、その後の学習の過程で単語のベクトル等が修正、変更されていく。

出力データは一つの特
定のテキストだけでな
く、他の大量のテキス
トの学習に影響される

おそらく、（技術の内容がわかっているやべきことをきちんとやっているという前提は必要だが）仮に学習データ中のテキストが丸ごとコピーされるとすると、同一のテキストをなん度も繰り返して学習する、あるいは偏って収集された非常に少量のデータだけで学習する等の条件が重なる必要があると推測

- 全く同一の学習データで学習しても、パラメータ数や計算の精度等が異なれば、全く異なるテキストが生成される

高齢者介護でコミュニケーションロボットが必要なのはなぜ？

40B

高齢者介護でコミュニケーションロボットが必要なのはなぜかという、高齢者は、認知症や脳梗塞などで、コミュニケーションがうまくとれないことがあるからです。コミュニケーションがうまくとれないと、介護する側もされる側も、ストレスがたまります。コミュニケーションロボットは、高齢者の話し相手になって、高齢者のストレスを軽減します。

認知症、脳梗塞
限定だが、スト
レス軽減が理由

179B, bf16

高齢者介護でコミュニケーションロボットが必要なのはなぜかという、高齢者は、介護者とのコミュニケーションを求めているからです。介護者が高齢者の話を聞いてあげることで、高齢者は安心します。また、高齢者は、介護者に話を聞いてもらうことで、自分の存在価値を確認することができます。

高齢者の「自己の
存在価値」まで踏
み込んでいるが、
対話の相手はなぜ
か介護者

179B, fp16

高齢者介護でコミュニケーションロボットが必要なのはなぜかという、高齢者の方は、自分の気持ちを表現するのが苦手な方が多いからです。自分の気持ちを表現するのが苦手な方は、自分の気持ちを表現するのが得意な方に比べて、ストレスを感じやすいということがわかっています。そのため、高齢者の方のストレスを軽減するために、コミュニケーションロボットが必要なのです。

自分の気持ちを
表現するのが苦
手な高齢者限定
だが、ストレス
軽減が理由



OpenAIのGPT-3並の1790億パラメータのモデル。350GBのWebテキストで学習
学習データはGPT-3よりかなり小さいがそれでも新聞1000年分の分量

NICT 大規模言語モデル (179B fp16)



トマトソースのスパゲティの作り方は以下の通り：



トマトソースのスパゲティの作り方は以下の通り：

1. フライパンにオリーブオイルを入れ、ニンニクを炒める。
2. ニンニクがきつね色になったら、トマトの水煮缶を入れ、トマトを潰しながら炒める。
3. トマトが煮崩れてきたら、塩、コショウ、バジルを入れ、味を調える。
4. 茹で上がったスパゲティを入れ、よく混ぜる。

応答文: "フライパンにオリーブオイルを入れ、ニンニクを炒める" に類似の学習データ中のテキスト

1. ".....フライパン...オリーブオイル....."
<http://banban.hatenablog.com/entries/2015/04/01>
2. ".....フライパン...オリーブオイル...ニンニク.....弱火.....香り....."
<http://www.nagoyatv.com/zakirobakeiko/oa/entry-3577.html>



トマトソースのスパゲティの作り方は以下の通り：

トマトソースのスパゲティの作り方は以下の通り：

1. フライパンにオリーブオイルを入れ、ニンニクを炒める。
2. ニンニクがきつね色になったら、トマトの水煮缶を入れ、トマトを潰しながら炒める。
3. トマトが煮崩れてきたら、塩、コショウ、バジルを入れ、味を調える。
4. 茹で上がったスパゲティを入れ、よく混ぜる。

応答文: "フライパンにオリーブオイルを入れ、ニンニクを炒める" に類似の学習データ中のテキスト

1. ".....フライパン...オリーブオイル....."
<http://banban.hatenablog.com/entries/2015/04/01>
2. ".....フライパン...オリーブオイル...ニンニク.....弱火.....香り....."
<http://www.nagoyatv.com/zakirobakeiko/oa/entry-3577.html>

- 学習データ中に類似テキストが存在する著作権侵害の恐れのあるテキストを生成した場合、赤下線で警告し、学習データ中のテキストの一部を表示
- 上記の実装では、著作権侵害に繋がりうるテキストの類似性を後述する「テキスト間の編集距離」で評価
- 注：検索エンジンに関する例外規定を使って学習データ中のWebテキストをURLとともに表示することは可能と推定→極端に長いテキストは除く。Bing、Bard等はどうか？

- 大規模言語モデルの構築に当たっては、様々な分野の情報からなる大量のWebテキストを学習。

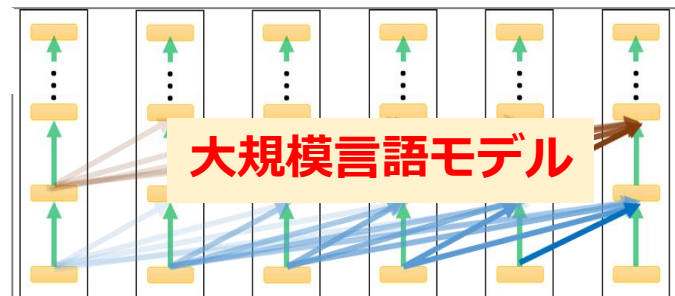
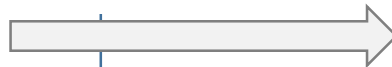
学習データ (大量のWebテキスト)



“...高齢者介護...介護モニタリング...人手不足...”
“フライパン...オリーブオイル...にんにく...弱火..
かおり”
“...本居宣長...江戸時代...古事記伝...伊勢...”

①

大量のWebテキストを学習して
大規模言語モデルを構築



- ①～④の流れで学習データと類似する出力テキストについて警告し、著作権侵害チェックを支援。

学習データ (大量のWebテキスト)

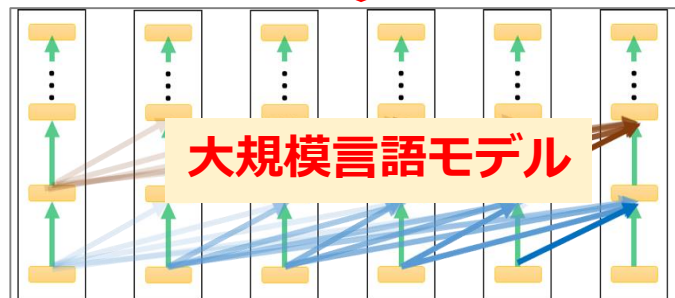


"...高齢者介護...介護モニタリング...人手不足..."
 "フライパン...オリーブオイル...にんにく...弱火...
 かおり"
 "...本居宣長...江戸時代...古事記伝...伊勢..."

(入力テキスト)

トマトソースのスパゲティの作り方は以下の通り：

- ① 大規模言語モデルにテキストを入力。



- ② 大規模言語モデルがテキストを出力。

1. フライパンにオリーブオイルを入れ、ニンニクを炒める。
2.

③

出力テキストと類似するテキストが学習データ中に含まれていないかを「テキスト間の編集距離」に基づき評価。

※今回の例では、出力テキスト（黄色）と学習データ（オレンジ）の類似性を検出

④

学習データ中に類似テキストが存在する出力テキストを赤下線で警告。合わせて、学習データ中の類似テキストの一部とURLを表示。

(出力テキスト)

トマトソースのスパゲティの作り方は以下の通り：

1. フライパンにオリーブオイルを入れ、ニンニクを炒める。
2. ニンニクがきつね色になったら、トマトの水煮缶を入れ、トマトを潰しながら炒める。
3. トマトが煮崩れてきたら、塩、コショウ、パジルを入れ、味を調える。
4. 茹で上がったスパゲティを入れ、よく混ぜる。

応答文: "フライパンにオリーブオイルを入れ、ニンニクを炒める" に類似の学習データ中のテキスト

1. ".....フライパン.....オリーブオイル....."
<http://banban.hatenablog.com/entries/2015/04/01>
2. ".....フライパン.....オリーブオイル.....ニンニク.....弱火.....香り....."
<http://www.nagoyatv.com/zakirobakeiko/oa/entry-3577.html>

- 著作権侵害チェック支援ツールにおいては、編集距離（レーベンシュタイン距離）に基づき、出力テキストと学習テキストの編集距離が一定以下の場合に類似しているものと評価
- 編集距離(レーベンシュタイン距離)

- レーベンシュタイン距離**（レーベンシュタインきょり、英: Levenshtein distance）は、二つの文字列がどの程度異なっているかを示す距離の一種である。**編集距離**（へんしゅうきょり、英: edit distance）とも呼ばれる。具体的には、1文字の挿入・削除・置換によって、一方の文字列をもう一方の文字列に変形するのに必要な手順の最小回数として定義される (Wikipedia「レーベンシュタイン距離」の項より転載。<https://ja.wikipedia.org/wiki/レーベンシュタイン距離> CC-BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0/deed.ja> に則り転載。注:文中の「一文字」は「一単語」でも良い)

例 テキスト1 : NICTでは大規模言語モデルを開発中です

↓
OpenAIでは大規模言語モデルを開発中です

↓
OpenAIでは大規模言語モデルを開発してです

↓
OpenAIでは大規模言語モデルを開発しています

↓
テキスト2 : 最近OpenAIでは大規模言語モデルを開発しています

テキスト1とテキスト2の
編集距離は4
(4回の単語置換・挿入を実施)

- 学習データにないテキストに関しては、それをLLMに出力しても依拠性がないため、著作権侵害にはならないと仮定

• 詳細設定をどう決めるか？

- 編集距離に関するしきい値はどう決める？
- 編集距離計算時に積算する単語置換と挿入は同じ重みでいいのか？
- 文単位で編集距離を計算する、でいいのか？ 複数個のテキストに分割したり、前後のテキストを考慮したりして、編集距離を計算すべきか？

今日は、文化庁審議会でのヒアリングで鳥澤が著作権侵害の検知を支援するツールの説明をし、編集距離についても説明したが、技術的課題も含めて課題はたくさんあるということについても述べた。(90文字)

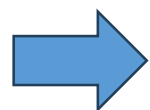
可能性1：この文まるまるを一つのテキストとして編集距離を計算

可能性2：この文まるまるは長いので、複数個のテキストに分割して編集距離を計算。しかし、分割のやり方は複数通りある。また、単純な分割でいいのか？「今日は、文化庁でのヒアリングで」は分割された複数テキストの先頭につけなくていいか？ また、前後のテキストは考慮しなくていいか？

- **創作性の認定が困難**：そもそも、テキスト表現に創作性が認められるか、あるいは例えば、単に客観的事実をシンプルに述べているだけなのかの判断は現状の技術で高信頼で判断できるか不明。おそらく困難
- **同義語の置き換えへの対応が困難**：
 - 彼は新聞のサイトの**スクショ**を自分のブログに載せました。
 - 彼は新聞のサイトの**コピー**を自分のブログに載せました。
 - 彼は新聞のサイトの**キャプチャー画像**を自分のブログに載せました。
 - これらの例では編集距離は大きくなってしまい、著作権侵害のリスクありと判定されないかもしれない
 - また、キャプチャー画像とコピーが同義語だと書いてあるような既存辞書はおそらく存在しない
- **翻案への対応が困難**：同義語ではないが、例えば、小説の主人公の名前を別の名前に置き換えた場合も同様のリスク
 - シャーロック・ホームズの小説の主人公の名前を全部フィリップ・マーロウに置き換えたら...
- 技術者的視点では、上記のような一定のルールに基づく対応が難しい課題を解決する現状最良の手段は、テキストの抽象的な表現にも対応できる大規模言語モデルを活用すること
 - 大規模言語モデルの生成テキストに係る課題がクリアにならないと大規模言語モデルによる課題解決も困難。鶏が先か、卵が先か？

- 仮にこうしたツールが有用だとして、以下の課題がある
 - **ユーザへの提示方法**：学習データ中の類似テキストと警告を提示するだけでいいか？
 - 企業が第三者にサービスとして提供する場合
 - **生成テキストの品質低下**：仮に学習データに類似テキストがあるものを出力せず、LLMに代替テキストを生成させるならば、以下が懸念される
 - テキストの品質低下
 - 無理筋のテキストを生成することによるハルシネーション、偽情報の増加
 - **運用コスト増大**：こうしたツールの導入を義務付けると、大量の学習データに関する検索エンジンの実装が必須となり、LLMの運用コストが増大
 - 特に国内ベンチャーによる商用化の障害となる可能性
 - （検索エンジン等に強みを持ち、巨大ITインフラを持つ）海外勢との競争で不利？
 - 国際的統一規制も重要

- 大規模言語モデルの学習プロセスを概観
 - 適切な大規模言語モデルの学習においては、仮に学習データ中のテキストが丸ごとコピーされて出力されたとしても、そのテキストの存在とコピーの間の因果関係は不明確
 - 学習アルゴリズムは基本全て数値計算で行われ、学習データ中の文字列の切り貼り等の操作は行わない
 - 特定の一つのテキストではなくて、大量の学習データ全体が生成テキストに影響する
 - 仮にテキストが丸ごとコピーされて出力されるとしたならば、様々な条件が重なったが故であると推測される
- 著作権侵害チェック支援ツール
 - NICTにおいて著作権侵害防止に資するツールを開発中
 - 技術的課題
 - その他の課題



引き続き、技術的課題や法的議論等を踏まえつつ、著作権侵害防止に資する研究開発を実施していく