

NICTにおける大規模言語モデルの試行構築について

- **NICT**では、従来より、**多言語翻訳技術**や**対話応答技術**などの、**AI自然言語処理の研究開発を実施**。
- 今般、**手持ちの計算機資源及び蓄積した日本語データを活用し、400億パラメータの日本語大規模言語モデルを試行構築**（2023年5月完成）。**7月4日、東大×生成AIシンポジウムにて松本大臣より発表**。同日、NICTより報道発表。
- 現行資源による、**学習の完走を優先した試行モデル**だが、**日本語でのやり取りが可能な水準に到達**しており、**今後の展開に手応え**（ただし、「短文単位での入出力」「洗練されていない日本語表現」など、性能面でChatGPT等と比較できるレベルにはまだない。）。
- **学習ノウハウ等については、国内研究者グループに提供済み**。今後、**広く国内のプレイヤーと連携すべく共同研究等を予定**。
- 更に、NICTでは**1790億パラメータ（GPT-3相当）の学習を実施中**（7月中旬頃学習完了見込み）。**更なる精度向上には、試行錯誤をしながら繰り返し学習を行う必要があります、計算機資源の一定規模の手当が必要**。（他に、良質な日本語データの拡充等も必要）

実際の入出力の様子（※学習の完走を優先し、短文単位での学習を行っているため、入出力も短文単位）

NICT 生成系大規模言語モデル(40B) デモ

The screenshot shows a chat interface with a blue character icon on the left. The user input is in a green box, and the model output is in a blue box. The output text is a detailed response about a cruise ship in the Aegean Sea. A red box highlights a specific phrase in the output: 「松阪市映画上映実行委員会」.

ユーザーの入力: エーゲ海での新婚旅行のプランは、

大規模言語モデルからの出力 (応答時間は数秒程度): エーゲ海での新婚旅行のプランは、エーゲ海の島々を巡るクルーズ船での旅が人気です。エーゲ海の島々は、ギリシャ本土から離れているため、クルーズ船での移動が便利です。クルーズ船での移動は、船内での食事やショーなども楽しめ、移動中も退屈しません。エーゲ海の島々では、ギリシャの伝統的な建物や遺跡などを見ることができます。また、エーゲ海の島々では、ギリシャの伝統的な料理やワインなども楽しめます。

ユーザーの入力: 本居宣長の映画のあらすじは

大規模言語モデルからの出力 (応答時間は数秒程度): 本居宣長の映画のあらすじは、国学者の本居宣長が、江戸時代の終わり頃に松阪の地で、国学の研究に没頭し、「古事記伝」を完成させるまでの生涯を描いたものです。松阪市では、この映画の上映をきっかけに、松阪の歴史や文化を再認識し、松阪の魅力を全国に発信するため、「**松阪市映画上映実行委員会**」を設立し、映画の上映を行います。この映画の上映にあたり、映画の上映を盛り上げるための企画を募集します。

NICT大規模言語モデル概要

- ◆ パラメータ数 : 397億
- ◆ 学習元データ: ウェブテキスト350GB (日本語web400億ページからフィルタリング・加工)
- ◆ 学習時間 : 計650.3時間(約27日)
- ◆ 学習に使用した計算機: NVIDIA A100 (80GB) 128枚
- ◆ 運用環境: NVIDIA A100 (80GB) 8枚

ユーザーの入力

大規模言語モデルからの出力
(応答時間は数秒程度)

架空の団体を創出?
(Googleでヒットせず)

✓ 一見もつと面白いが、誤りや架空の回答も出力。いわゆる「ハルシネーション」だが、創造性の裏返し

✓ NICTでは、創造性を残しつつ、生成する文章からハルシネーションや、著作権・プライバシー等の権利侵害に当たる内容を検出・除去するような、「安心して使える」技術も併せて研究中

日本語に特化した大規模言語モデル(生成 AI)を試作

～日本語の Web データのみで学習した 400 億パラメータの生成系大規模言語モデルを開発～

【ポイント】

- 400 億パラメータの大規模言語モデル(生成 AI)を Web から収集した 350 GB の日本語テキストを用いて開発
- 従来から利用していた高品質な日本語テキストを事前学習することで、約 4 か月で開発
- 現在 1,790 億パラメータのモデルも学習中。今後は民間企業、国研、大学等と共同研究等を通して更に発展

国立研究開発法人情報通信研究機構(NICT、理事長: 徳田 英幸)は、ユニバーサルコミュニケーション研究所データ駆動知能システム研究センターにおいて、独自に収集した 350 GB の日本語 Web テキストのみを用いて 400 億パラメータ¹の生成系²の大規模言語モデル³を開発しました。今回の開発を通し、事前学習⁴用テキストの整形、フィルタリング、大規模計算基盤を用いた事前学習等、生成系の大規模言語モデル開発における多くの知見を得ました。現在は、更に大規模な 1,790 億パラメータの生成系大規模言語モデル(OpenAI 社の GPT-3⁵と同等規模)の学習を実施中で、また、学習用テキストの大規模化にも取り組んでいます。今後、共同研究等を通して民間企業、国研、大学等と協力して、日本語の大規模言語モデルの研究開発や利活用に取り組む予定です。

【背景】

NICT では、これまで Web ページを収集し、インターネット上にある知識を有効活用するため、深層学習⁶を自然言語処理技術に応用し、様々な研究開発に取り組んできました。2018 年に識別系²の言語モデルと呼ばれる BERT⁷が発表されてから、独自に収集した日本語の Web テキストを用いて、BERT を基に日本語用に改良した識別系言語モデルを構築し、2021 年に試験公開を開始した大規模 Web 情報分析システム WISDOM X 深層学習版⁸や高齢者介護支援用対話システム MICSUS⁹等で活用してきました。そして、識別系言語モデルとしては大規模な、200 億パラメータのモデルを構築するなど、大規模言語モデルの構築に関するノウハウを蓄積してきました。

一方、生成系言語モデルは、2018 年に発表された GPT を始めとして、様々な研究が行われてきましたが、生成系言語モデルには、偽情報やバイアス、各種悪用のリスクが指摘されており、NICT では、こうしたリスク解決のために時間を要し、その活用は少し先になると考えていたため、生成系言語モデルの研究開発を基礎研究と位置付け、これまで大規模には実施してきませんでした。しかし、2022 年の ChatGPT の登場以降、生成系言語モデルの有効性に大きな注目が集まり、その開発力が他分野にも影響されると指摘される中、我が国における開発力強化が喫緊の課題となっていることから、NICT としても、生成系言語モデルの開発に本格的に取り組み始めたものです。

【今回の成果】

この度、NICT は、これまで構築してきた識別系言語モデルの学習に用いていた、ノイズに相当するテキストが少ない 350 GB の高品質な独自の日本語 Web テキストを用いて、400 億パラメータの生成系の大規模言語モデルを事前学習し、その動作の検証を開始しました。これまでのノウハウの蓄積もあり、ユーザインタフェースを含め 4 か月程度で完了しました。今回は、学習の完了を優先させて事前学習を実施しており、ファインチューニング¹⁰や強化学習¹¹は未実施であるため、短めの入出力、洗練されていない日本語表現など、性能面では ChatGPT 等と比較できるレベルではありませんが、図 1 に示すように、日本語でのやり取りが可能な水準に到達しています。

具体的には、要領を得ないテキストが出力されるケースも多々あるものの、補足資料に示すように、各種質問への回答、要約、論文要旨の生成、翻訳などが可能になっています。加えて、存在しない映画の簡単なあらすじを生成す

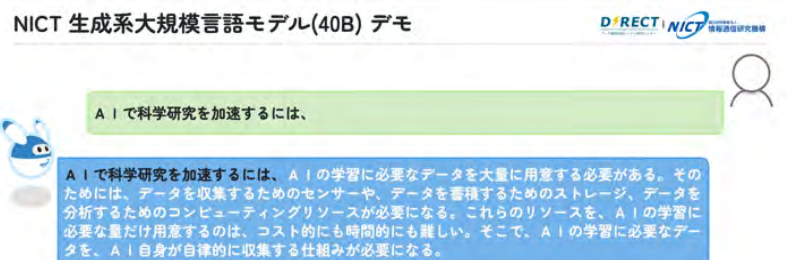


図 1: NICT の大規模言語モデルプロトタイプの実動作例

具体的には、要領を得ないテキストが出力されるケースも多々あるものの、補足資料に示すように、各種質問への回答、要約、論文要旨の生成、翻訳などが可能になっています。加えて、存在しない映画の簡単なあらすじを生成す

るといった一種の創作ができる可能性も示しています。一方で、生成テキストの悪用の可能性を示唆する結果も得られており、今後、ポジティブ、ネガティブの両方の要素に関して改善を図っていく予定です。また、著作権侵害の問題に関しては、生成したテキストに類似するテキストが学習データにないかを自動検索し、著作権侵害のチェックを容易にしています。

今後、NICT でこれまでに蓄積してきた人手により作成した大量の学習データ、例えば、WISDOM X の 150 万件を超える質問応答用データ等を活用して、ファインチューニング等を行い、品質を高め、具体的なアプリケーションでの活用を容易にしていく予定です。

【今後の展望】

今回の成果は、400 億パラメータという生成系の日本語大規模言語モデルを試作(事前学習)したというものが、事前学習に用いるテキストが十分であるとは考えていません。今後は、学習用のテキストについて、日本語を中心として更に大規模化していきます。また、現在、GPT-3 と同規模の 1,790 億パラメータのモデルの事前学習に取り組んでおり、適切な学習の設定等を探索していく予定です。さらに、より大規模な事前学習用データ、大規模な言語モデルの構築に際し、既に述べたポジティブ、ネガティブの両方の要素に関して改善を図るとともに、WISDOM X、MICSUS 等既存のアプリケーションやシステムの高度化等に取り組む予定です。加えて、NICT では、まだ誰も考えておらず、Web 等にも書かれていない、具体的で「尖った」将来シナリオや仮説をテキストとして生成し、対話システムによるブレインストーミング等で活用するための研究を実施してきましたが、このような研究においても今回開発した日本語大規模言語モデル等を活用していく予定です。さらに、民間企業、大学、国研等との共同研究等を拡大し、安全に活用可能な大規模言語モデルの研究開発等に産学官連携して取り組む予定です。共同研究等に関する問合せは、以下の〈本件に関する問合せ先〉までお願いいたします。

〈関連する過去の NICT の報道発表等〉

- ・2021 年 3 月 31 日 大規模 Web 情報分析システム WISDOM X「深層学習版」の試験公開を開始
<https://www.nict.go.jp/press/2021/03/31-3.html>
- ・2023 年 3 月 8 日 高齢者向け対話 AI でケアマネジャー面談業務時間の 7 割削減に成功
<https://www.nict.go.jp/publicity/topics/2023/03/08-1.html>

〈 本件に関する問合せ先 〉

国立研究開発法人情報通信研究機構
ユニバーサルコミュニケーション研究所
データ駆動知能システム研究センター
大竹 清敬
E-mail: wisdom-contact@ml.nict.go.jp

〈 広報（取材受付） 〉

広報部 報道室
E-mail: publicity@nict.go.jp

<用語解説・補足説明>

*1 パラメータ

ニューラルネットワークは、人間の脳にある脳細胞のネットワークをモデル化したものである。基本的に、このネットワークにおける結線一つ一つには「重み」という数値が持たされており、この重みをパラメータという。例えば、1億パラメータのニューラルネットワークとは、こうした重みが1億個あることを意味する。最近では、パラメータの数は、ニューラルネットワークの規模を表す指標として用いられている。

*2 生成系と識別系の言語モデル

ここでは、日本語や英語といった自然言語のテキストやプログラム等の長い文字列を生成できる言語モデルを生成系言語モデルと呼び、人工的な分類ラベルを少数出力する言語モデルを識別系言語モデルと呼んでいる。ChatGPT や今回 NICT で構築した言語モデルは生成系言語モデルである。一方、文中で参照されている BERT は、通常、人工的な分類ラベルを数個出力するだけなので、識別系言語モデルである。この人工的な分類ラベルは、質問応答や文書分類といったタスクごとに変わってよく、例えば、YES/NO、1/0 のように単純な二値のラベルもあれば、より多数の値のラベルを使うこともある。また、YES/NO という二値のラベルのいずれかを出力する単純なモデルであっても、例えば、質問(例: 富士山の高さは?)と回答の候補(例: 2,000メートル)のペアを入力として、回答の候補が質問に対する正解であれば YES、不正解であれば NO と出力するように学習を行えば、質問応答システムでの活用が可能になる。実際、NICT が試験公開中である質問応答システム WISDOM X は、BERT をそのような形で学習させ、活用している。なお、識別系言語モデルは、あらかじめ設定された少数のラベルを出力するだけなので、想定外の出力がなされることは少なく、偽情報やバイアスといったリスクは生成系言語モデルに比べて小さい。(なお、識別系言語モデルでも、前述のようなリスクのある使い方は考えられる。)また、同様のラベルを出力させることは、生成系言語モデルでも可能である。例えば、「富士山の高さは?」という質問と「2,000メートル」という回答候補のペアを入力としたときに、「回答候補は不正解」というテキストを出力させる一方、「富士山の高さは?」と「3,776メートル」のペアを入力としたときには、「回答候補は正解」と出力させるように学習させれば、生成系言語モデルを識別系言語モデルのように使うことも可能である。また逆に、識別系言語モデルも追加で学習をさせることで、生成系言語モデルに変換することも可能であり、NICT でもそのような使い方を過去に実施している。つまり、一定の作業を行えば、両者は互いに変換可能であるが、最終的に高性能が得られるかどうかは実際に変換してみなければ分からず、詳細に検討する必要がある。

*3 大規模言語モデル(Large Language Model; LLM)

非常に巨大な学習データ(テキスト)と深層学習を用いて構築された巨大なニューラルネットワークによる言語モデルである。言語モデルとは、基本的には単語の並びが与えられると、それら単語の並びが出現する確率を計算するものであるが、GPT 等の大規模言語モデルは、与えられたテキストに対し後続する単語を確率的に予測し、確率が最大となる単語を出力することを繰り返し、それらの単語を繋いでいくことでテキストを生成することができる。

*4 事前学習(Pre-training)

事前学習とは、深層学習による大規模言語モデル等の構築において、大量のテキストの一部をランダムに選んでニューラルネットワークからは見えないようにし、単語の穴埋め問題を大量に作ってそれを解くことをタスクとする学習(BERT)や、学習テキストの一部をランダムに選んで、その後続の単語を予測するというタスクを設定し(GPT等)、モデルを学習することである。重要な点は、これらの学習に際しては、Web ページのような人が書いた大量のテキストさえあれば、従来のように人が別途ラベル付けや作文をしなくても、ランダムに単語やテキストの一部を選ぶだけで、大量の学習データを自動で作ることができ、少なくとも学習データの確保という点では極めてローコストになるということである。大規模言語モデルの登場以前は、目標とするタスクに関して人がラベル付け、作文等を行った学習データで、ニューラルネットワークを学習させ、そのタスクを実施していたが、この人によるラベル付け等、学習データ作成の作業コストの高さが課題となっていた。一方、大規模言語モデルの登場以後、作成コストが高い人手作成の学習データが少量しかなくても、事前学習済みの大規模言語モデルをその少量のデータで追加学習(ファインチューニング)させることで、高い精度を出せることが示された。この結果、事前学習済みの大規模言語モデルを一つ用意しておけば、様々なタスクに特化した比較的少量の人手作成学習データでファインチューニングさせることで、学習データ作成のコストを抑制しつつ多数のタスクで高精度を実現できるようになった。また、このように事前学習済みの大規模言語モデルを「使い回す」ことが一般的になった状況もあって、事前学習済み大規模言語モデルは、基盤モデルとも呼ばれるようになった。ただし、現在は、基盤モデルという用語は、大規模言語モデルのみならず、より幅広いタスクに適用可能な大規模人工知能モデルといった意味で用いられる。

*5 GPT(Generative Pre-trained Transformer)

GPTは、OpenAI が2018年に発表した、Transformer^{*12}ベースの言語モデルで、大規模なテキストコーパスで学習し、人間のようなテキストを生成することが可能なモデルである。2018年に発表されたモデルは1億1,700万パラメータだったが、2019年に15億パラメータのGPT-2、2020年に1,750億パラメータのGPT-3が発表され、大規模化が進んできている。OpenAIは、2022年にGPT-3を発展させたGPT-3.5を用いてファインチューニングや強化学習を行い、

人間のようにチャットすることが可能な AI として ChatGPT を公開した。その後、2023 年 3 月に GPT-4 を ChatGPT に導入するなど、更に発展させているが、GPT-3.5 や GPT-4 のパラメータサイズ等に関する詳細は、公式には明らかにされていない。

*6 深層学習 (Deep Learning; ディープラーニング)

深層学習とは、従来、学習が困難だった複雑なニューラルネットワークを学習するための機械学習手法や、各種の複雑なニューラルネットワークアーキテクチャ等の総称である。深層学習の発展により、Transformer を始めとする複雑なニューラルネットワークのアーキテクチャや、それをういた大規模言語モデル等が可能となった。

*7 BERT (Bidirectional Encoder Representations from Transformers)

BERT は、Google が 2018 年に発表した、Transformer ベースの識別系大規模言語モデルで、文書分類や同義判定等のタスクに向いている。2018 年に Google が発表した英語版のモデルは 3.4 億パラメータである。

*8 大規模 Web 情報分析システム WISDOM X (ウィズダム エックス) 深層学習版

WISDOM X は、大規模な Web ページを対象とした質問応答システムとして、2015 年から NICT が試験公開しているシステムである。なぜ、何、どうやって、どうなるといった多様な質問に回答することが可能で、2021 年 4 月からは、BERT を取り入れた深層学習版として試験公開を継続中である。例えば、「AI ってどんな社会課題の解決に使えるのかな」「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」「量子コンピュータが実用化されるとどうなる」といった多様な質問の回答を Web から網羅的に抽出し、端的に提示することができる。

<https://www.wisdom-nict.jp/> にて、どなたでも活用できる。

*9 高齢者介護支援用対話システム MICSUS (ミクス)

MICSUS は、要支援等の認定を受けている在宅高齢者に対して、介護職であるケアマネジャーが実施する健康状態や生活習慣のチェックを行う介護モニタリングと呼ばれる作業の一部を代替し、ケアマネジャーの作業負担を軽減することを狙って内閣府の SIP 第 2 期 (2018 年度～2022 年度) の支援の下、KDDI 株式会社、NEC ソリューションイノベータ株式会社、株式会社日本総合研究所と NICT が共同開発したシステムである。WISDOM X の技術を応用し、Web 情報に基づいた雑談も可能とすることで、飽きることなく介護モニタリングを実施し、普段使いしてもらうことで、現在は最低、月に 1 度となっている介護モニタリングの頻度向上による介護の質の改善を目指すとともに、高齢者の健康状態悪化の要因の一つといわれているコミュニケーション不足の解消も狙ったシステムである。現在は、KDDI が NICT、NEC ソリューションイノベータ、日本総合研究所と協力して、社会実装に向けて多様なパートナー企業と共同実証や事業化を検討している。

https://direct.nict.go.jp/news/#MICSUS_NEWS_20230308 にて、これまでの実証実験の結果や実証実験の様子をご覧いただける。

*10 ファインチューニング

事前学習済みのモデルを別のデータで追加で学習することをファインチューニングと呼ぶ。特に、自然言語処理分野では、事前学習済みの大規模言語モデルを、目的とするタスクに特化した学習データでファインチューニングすることで、高性能を得ることができることが知られている。

*11 強化学習

強化学習とは、ある環境におけるエージェント (例えば、対話システム) が、環境とのやり取り (例えば、対話) を通して得られる報酬を最大化する行動を学習していく機械学習技術である。例えば、ChatGPT では、モデルをファインチューニングした後に、実際に人間との対話を通して、強化学習を行っているといわれている。

*12 Transformer

Transformer は、2017 年に Google が発表したニューラルネットワークのアーキテクチャで、当初は、機械翻訳を目的として開発されたが、BERT や GPT で使われるようになり、機械翻訳以外の自然言語処理へ応用され、さらには、画像処理等にも応用が広がっている。

今回開発した 400 億パラメータの大規模言語モデル(以下「日本語大規模言語モデル」)は、独自に収集した 350 GB の日本語 Web テキストで事前学習が完了した状態で、ファインチューニング、強化学習等は実施していません。用意したユーザインタフェースでは、入力したテキスト(プロンプト)に対し、日本語大規模言語モデルがその後続のテキストを生成します。入力を工夫することで、現在の状態でも様々なタスクを実施できる可能性を感じ取ることができると同時に、事実と異なる内容を含むテキストや意味をなさないテキストを生成することも確認しています。

NICT 生成系大規模言語モデル(40B) デモ



Example 1:
 User: エーゲ海での新婚旅行のプランは、
 AI: エーゲ海での新婚旅行のプランは、エーゲ海の島々を巡るクルーズ船での旅が人気です。エーゲ海の島々は、ギリシャ本土から離れているため、クルーズ船での移動が便利です。クルーズ船での移動は、船内での食事やショーなども楽しめ、移動中も退屈しません。エーゲ海の島々では、ギリシャの伝統的な建物や遺跡などを見ることができます。また、エーゲ海の島々では、ギリシャの伝統的な料理やワインなども楽しめます。

Example 2:
 User: AI で科学研究を加速するには、
 AI: AI で科学研究を加速するには、AI の学習に必要なデータを大量に用意する必要がある。そのためには、データを収集するためのセンサーや、データを蓄積するためのストレージ、データを分析するためのコンピューティングリソースが必要になる。これらのリソースを、AI の学習に必要な量だけ用意するのは、コスト的にも時間的にも難しい。そこで、AI の学習に必要なデータを、AI 自身が自律的に収集する仕組みが必要になる。

Example 3:
 User: トマトソースのスパゲティの作り方は以下の通り：
 AI: **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. ニンニクの香りが出てきたら、タマネギを加え、タマネギがしんなりするまで炒める。
 3. トマトの水煮を加え、トマトを潰しながら炒める。
 4. 塩、コショウで味を調え、バジルを加える。
 5. 茹で上がったスパゲティを加え、よく混ぜる。
 6. 皿に盛り、バジルを飾る。
 応答文: "ニンニクの香りが出てきたら、タマネギを加え、タマネギがしんなりするまで炒める" に類似の学習データ中のテキスト
 1. ".....ニンニク...ピッキーマス...ざく切り...タマネギ...肉.....タマネギ.....する...炒める....."
<http://mhshine.net/2009/11/>

図 2: NICT の日本語大規模言語モデルプロトタイプの実動作例(1)

図 2 に、今回の日本語大規模言語モデルが、ユーザが与えたテキストの続きを生成した動作例を示しています。ユーザーアイコンからの緑の吹き出しに入力を、NICT のキャラクター、人工知能“N”からの青色の吹き出しに日本語大規模言語モデルからの出力を表示しています。白でハイライトされている部分が、日本語大規模言語モデルが生成したテキストになります。

1 つ目の例は、たまたま、我々の関係者でエーゲ海に新婚旅行に行った者がいたため、出力させてみたものですが、この生成したテキストの内容どおりに、クルーズ船での新婚旅行を楽しんだとのことでした。なお、この入力日本語として整ってはいませんが、大規模言語モデルに与える質問であると捉えることもできます。今後、ファインチューニング等によって「エーゲ海でのおすすめの新婚旅行のプランを教えてください。」等、より整った形の質問にも対応可能としていきます。

2 つ目の例は、AI で科学研究を加速するには、AI 自身が自律的に学習に必要なデータを収集する仕組みが必要になり、それによって各種リソースの無駄が省けるかもしれない、という示唆に富んだ内容を示しています。

3 つ目の例は、料理のレシピが箇条書きで出力された例ですが、現在、我々が開発中の、著作権侵害の検出を容易にする自動検索ツールが、生成されたテキストと類似する文を学習データ中で見つけた場合の動作例です。生成した文と類似する文が学習データに見つかった際は、赤い点線の下線でそれを知らせ、該当する文をクリックすることで、検索結果を生成文の下に出力し、そのテキストが収集された URL と共に確認できるようになっています。なお、学習データ中の類似テキストをそのまま表示すると、そのこと自体が著作権侵害となる恐れがあることから、この図では中心的なキーワードのみを示しています。

本居宣長の映画のあらすじは

本居宣長の映画のあらすじは、国学者の本居宣長が、江戸時代の終わり頃に松阪の地で、国学の研究に没頭し、「古事記伝」を完成させるまでの生涯を描いたものです。松阪市では、この映画の上映をきっかけに、松阪の歴史や文化を再認識し、松阪の魅力を全国に発信するため、「松阪市映画上映実行委員会」を設立し、映画の上映を行います。この映画の上映にあたり、映画の上映を盛り上げるための企画を募集します。

奈良の鹿は、毎年6月ごろに赤ちゃんが生まれるんですよね。私も以前、ニュースで見て、実際に奈良に見に行ったことがあります。鹿の赤ちゃんは、ほんとに小さくてよちよちとお母さんを追いかけていたりして可愛かったです。他の大人の鹿たちも赤ちゃんの様子を見守っていて、心が温まりました。でも、あまり赤ちゃんに近づくとお母さんが怒って攻撃することがあるらしく、注意が必要だそうです。ご参考までに。／要点：

奈良の鹿は、毎年6月ごろに赤ちゃんが生まれるんですよね。私も以前、ニュースで見て、実際に奈良に見に行ったことがあります。鹿の赤ちゃんは、ほんとに小さくてよちよちとお母さんを追いかけていたりして可愛かったです。他の大人の鹿たちも赤ちゃんの様子を見守っていて、心が温まりました。でも、あまり赤ちゃんに近づくとお母さんが怒って攻撃することがあるらしく、注意が必要だそうです。ご参考までに。／要点：奈良の鹿は、毎年6月ごろに赤ちゃんが生まれるので、その時期に奈良に行くと赤ちゃん鹿を見ることができます。

健康でストレスなく暮らすコツを3つ挙げると

健康でストレスなく暮らすコツを3つ挙げるとしたら、

1. 自分の体を知ること
2. 自分の体に合った食事をすること
3. 自分の体に合った運動をすることです。

健康でストレスなく暮らすコツを4つ挙げると

健康でストレスなく暮らすコツを4つ挙げるとすれば、

1. 自分の体を知ること
2. 自分の体に合った食事をすること
3. 自分の体に合った運動をすること
4. 自分の体に合った睡眠をとることです。

図 3: NICT の日本語大規模言語モデルプロトタイプの動作例(2)

図 3 にその他の例を示します。1 つ目の例は、存在しない映画のあらすじや、その映画にまつわるプロモーション活動に関するテキストが創作されている例です。我々が調べた範囲では、本居宣長に関する映画も「松阪市映画上映実行委員会」も実在しません。いわゆるハルシネーションの例ですが、一種のクリエイティビティと考えることもできると思います。2 つ目の例は、比較的長いテキストを要約させたものです。なお、この要約の例の入力は、ブログに似せて我々が作文したテキストです。3 つ目、4 つ目の例は、3 つ、4 つと数を正しく認識して、指定された数の箇条書きを生成できている例になります。