

東北大学萩友会九州交流会「東北大学フォーラム2024 in 福岡」

文章生成AIとは何者か？

—動作原理，得意/不得意なこと，なにをもたらすのか—

鈴木潤

東北大学 言語AI研究センター
センター長・教授

- 導入：文章生成AI とは？
- 文章生成AI の成り立ち
- 文章生成AI の特性
- 文章生成AI の展望
- まとめ

[注意事項] 事前のお断り

- 本資料の情報の正確性について
 - 文章生成AI 関連の情報は**頻繁に更新**
 - あくまで**2024.01.28**時点の情報

本資料の情報が時間経過と共に正しくなくなる可能性があることに注意

自己紹介

- 鈴木 潤 東北大学 / 教授
 - **言語AI研究センター センター長** (2023.10.01 新設)
 - 大学院情報科学研究科 **人工知能基礎学講座** (協力講座)
 - 他複数の部局を兼務

職歴

- 2001.04-2018.03 **日本電信電話株式会社** コミュニケーション科学基礎研究所
- 2018.04-2020.06 東北大学 大学院情報科学研究科 乾・鈴木研究室 准教授
- 2020.04-2022.04 **Google LLC** Visiting Researcher (クロスアポイントメント)
- 2020.07-2023.09 東北大学 データ駆動科学・AI教育研究センター (研究室は2021.4から)
- 2023.10- 現在 現職

- 専門分野：**人工知能, 自然言語処理, 機械学習**
 - **計算機が言語を効率的に学習する方法の研究に従事**

- 異なる部局に所属する研究者を結集し、AI基盤研究からAI共生社会構築/社会実装までを推進する世界的研究拠点を旨す
- 日本の大学で初めての言語AIに特化した研究センター

言語AI研究センター - 自然言語処理/生成系AI研究の新展開 -

【基盤研究部門】

生成系AIを含む
高度AIの信頼できる
AI (Trustworthy AI) 化

【領域適応研究部門】

各専門領域における
AI駆動型研究

【AI共生社会研究部門】

国内外と連携した
AIアラインメント/セイフティ研究

【社会共創部門】

研究成果の社会実装と
人材育成



- 導入：文章生成AIとは？
- 文章生成AIの成り立ち
- 文章生成AIの特性
- 文章生成AIの展望
- まとめ

例：ChatGPT

<https://chat.openai.com/>



ChatGPT 3.5 ▾



今日は何をお手伝いしましょうか？

データベースのスキーマを設計する
オンライングッズストア用

メールを書く
私のプロジェクトの締め切りを延長...

ChatGPTにメッセージを送る...



ChatGPTは間違いを犯すことがあります。重要な情報は確認をお考えください。

文章生成AIとは何者か？ - 動作原理, 得意/不得意なこと, なにをもたらすのか - / 2024.01.28

6

文章生成AI / ChatGPT とは？

- 一言で言うと...

[一般向け]

(人が人に話すような) **対話形式の指示**を受け付け
その指示に適した文章を生成する **文章生成器**

(それ以上でもそれ以下でもない)

[自然言語処理分野の研究者/技術者向け]

大規模データで事前学習し **指示文**で微調整学習した
大規模言語モデル (Large Language Model: LLM)

文章生成AIとは何者か？ - 動作原理, 得意/不得意なこと, なにをもたらすのか - / 2024.01.28

7

利用例

- 論文作成支援
- 文章要約
- 翻訳
- 文章スタイル変換
- レポート作成
- 目次生成
- ブレインストーミング
- 旅行プラン作成
- 料理メニュー作成
- (ゲーム)シナリオ生成
- ギャグの生成/分析
- スピーチ作成
- 語学学習支援
- 面接練習
- プログラミング生成/支援
- アプリ開発支援
- カウンセリング
- カスタマーサポート
- キャリア診断
- 法律相談
- 演習/レポート課題作成
- . . .

人間の様々な知的活動をサポート可能

=> 使い方の工夫次第で無限の可能性!?

[注意点] ChatGPT = (対話型)文章生成AI ?

- ChatGPT
 - ✕ 対話にて利用者の要求に応えるシステムの総称?
=> 「(対話型)文章生成AI」または「言語モデル」
 - ○ OpenAIが構築した(言語モデルの)モデル名
 - 「GPT-4」なども同様にモデル名
- 誤用の原因
 - 報道等で「文章生成AI」と呼ばずに「ChatGPT」で統一して利用
=> モデル名が類似システムの総称に
類似例! iPod = 携帯型音楽プレイヤー (若者向けは iPod = 携帯型音楽プレイヤーらしい)

まとめ：文章生成AIとは？

- 文章生成AI を一言で説明すると？

(人が人に話すような) 対話形式の指示を受け付け
その指示に適した文章を生成する文章生成器

- 人間の様々な知的活動をサポート可能
=> 使い方の工夫次第で無限の可能性 !?
- ChatGPT = (対話型)文章生成AI ?

- 導入：文章生成AI とは？
- 文章生成AI の成り立ち
- 文章生成AI の特性
- 文章生成AI の注意点と今後の展望
- まとめ

文章生成AI を構成する技術：まとめ図

2階部分 (微調整学習)

- 指示文設計
- 指示文と適切な回答データの学習



- 獲得するスキル
- 指示文の理解
 - 対話的なやりとりの理解
 - 不適切発言の抑制

1階部分 (事前学習)

- 大規模言語ニューラルネットワーク
- 大規模文章データの丸覚え学習



- 獲得するスキル
- 文章の流暢さ
 - 世界の知識

土台 (言語モデル)



基盤：(従来の) 言語モデル

- 文章の出現確率を予測する **確率モデル**

語彙
(単語の集合)

に
。
これ
です
今日
1
・・・
仙台
良い
東京
は
・・・
天気
・・・

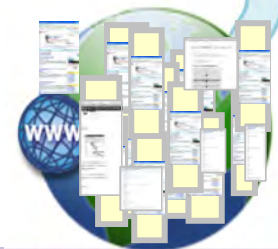
$P(Y = \$BOS\$ \text{今日は良い天気です。}\$EOS\$)$ 確率 高

$P(Y = \$BOS\$ \text{良いです。天気は今日}\$EOS\$)$ 確率 低

文章の出現確率は大量のデータから何かしらの形で計算

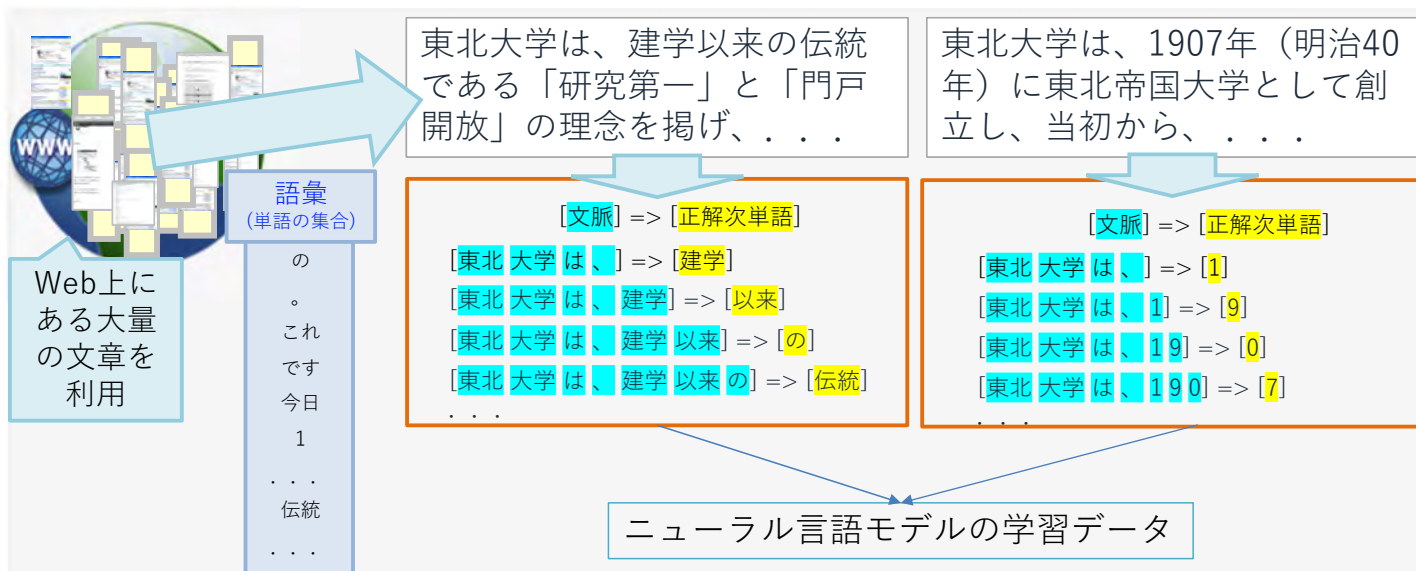
例：Web上にある文章中に出てきた文字列の頻度

言語モデル



「最近の」言語モデル

- 前提：ニューラル言語モデル
 - 文章の出現確率を深層ニューラルネットワークで近似
 - [文脈] => [正解次単語] を分類問題として学習



「最近の」言語モデル

- 前提：ニューラル言語モデル
 - 文章の出現確率を深層ニューラルネットワークで近似
 - [文脈] => [正解次単語] を分類問題として学習
 - 前方の単語列 (文脈) $\mathbf{Y}_{<j}$ を利用し次の単語 \hat{y}_j を予測 (繰り返す)

$$\hat{y}_j \sim P_{\theta}(y_j | \mathbf{Y}_{<j})$$

「最近の」言語モデル

- 前提：ニューラル言語モデル

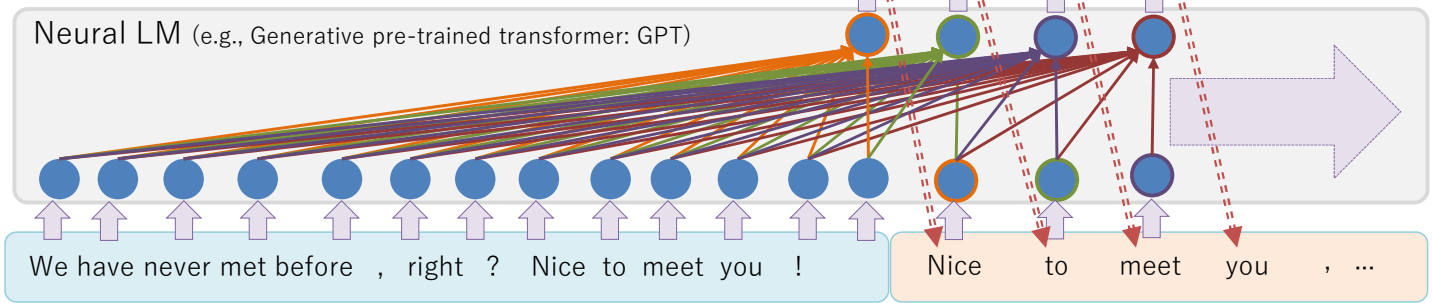
- 文章の出現確率を深層ニューラルネットワークで近似
- [文脈] => [正解次単語] を分類問題として解く
- 前方の単語列 (文脈) $Y_{<j}$ を利用して (繰り返す)

$$\hat{y}_j \sim P_{\theta}(y_j | Y_{<j})$$

Vocabulary	Vocabulary	Vocabulary	Vocabulary
A	A	A	A
this	this	this	this
that	that	that	that
...
meet	meet	meet	meet
have	have	have	have
you	you	you	you
...
Nice	Nice	Nice	Nice
...
to	to	to	to
...
too	too	too	too
,	,	,	,
.	.	.	.

OUTPUT text ~>

Nice to meet you , too ...



文章生成AI の成功要因は？

- **×** 言語モデル？
 - 言語モデルはただの確率モデル
 - 単体では大きな効果はえられない
 - 出現頻度を大きなデータから計測できてもChatGPTにはならない
- **○** DNNで文章の出現確率を近似
 - 言語 (記号) の分散表現 => 類似度が計算可能に
 - 大規模なデータを丸覚え「記憶」
- **○** 記憶の「呼出」を制御する指示文 (プロンプト) 設計
 - プロンプト = 文脈
 - プロンプトの与え方で自在に出力 (生成文) を制御

文章生成AI を構成する技術：まとめ図

2階部分 (微調整学習)

- 指示文設計
- 指示文と適切な回答データの学習



獲得するスキル
指示文の理解
対話的なやりとりの理解
不適切発言の抑制

1階部分 (事前学習)

- 大規模言語ニューラルネットワーク
- 大規模文章データの丸覚え学習



獲得するスキル
文章の流暢さ
世界の知識

土台 (言語モデル)



指示文(プロンプト)設計



まとめ：文章生成AIの成り立ち

- 土台：言語モデル
 - 文章の出現確率を予測する確率モデル
 - ニューラル言語モデル：文章の出現確率を深層ニューラルネットワークで近似
- 言語モデルの事前学習と微調整学習により実現
 - 事前学習：記憶
 - 微調整学習：記憶の呼出
- 指示文(プロンプト)設計により様々な出力が可能
 - 適切な指示文(=呼出)の設計は難しい/試行錯誤が必要

- 導入：文章生成AIとは？
- 文章生成AIの成り立ち
- 文章生成AIの特性
- 文章生成AIの展望
- まとめ

[注意事項] ChatGPTに関する情報

GPT-4 Technical Report

<https://cdn.openai.com/papers/gpt-4.pdf>

詳細不明 !!

[簡単な訳] GPT-4のような大規模モデルの競争環境と安全性を考慮し本報告書では、アーキテクチャ(モデルサイズを含む)、計算環境、学習計算量、データ、学習方法などに関する詳細な情報を記載しない。

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. **Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.**

ChatGPT (OpenAI) や Bard (Google) も同様に詳細は公表しない方針

今後、大手IT企業がサービスとして利用する対話型文章生成AIの詳細な情報は出てこない見込み

ChatGPTが注目された理由

- ユーザーのあらゆる要求に対応(しようと頑張る)
 - 例: 質問応答, 翻訳, 文書要約, 文章構成, スピーチ/レポート作成, ブレスト相手, 語学学習, プログラム生成, ...
- ⇒ (全ての要求に適切に回答できるわけではないが) そこそこ妥当な回答
- 相手が人間である場合とほぼ同じ感覚で自然言語の文章で要求することができる
 - 「対話形式の指示」により実現
 - 専門知識が不要
- 一般に普及するのはずっと先とされていたクオリティのサービスが無料でいきなり提供

既知の技術的課題

- 情報の正確性 (Hallucination)
 - 自信をもって間違いを提示
 - 現在情報の真偽を判断する機構がない
- バイアス (Bias)
 - 特定の偏った意見排除の困難性
 - 特定の個人や団体の利益になる応答を出すように調整可能
- 有害情報の取得
 - (悪意あるユーザーの) 特殊な指示により有害情報が提示可能
- プライバシー/コンプライアンス保護の困難性
 - 入力したデータは学習に利用される可能性 (ただしopt-out可能)
 - ユーザのリテラシレベルの向上が必須かつ急務
- 生成文の元情報を辿ることが困難
 - 著作権問題の主要因
- 技術の独占/民主化
 - 大手IT企業による間接的な支配
 - 公平性/信頼性の欠如に繋がる

課題例：誤情報の提示

j バレーボールで回転レシーブを考案したのは誰ですか？

バレーボールで回転レシーブを考案したのは、日本の元バレーボール選手である石井嘉幸 (いしいよしゆき) 氏です。石井嘉幸は、1990年代に日本の男子バレーボール代表として活躍し、特にその優れたレシーブ技術で知られていました。彼は独自のスタイルである回転レシーブを開発し、これは後に多くのプレーヤーやチームに影響を与えました。その後、バレーボール界での彼の功績が認められ、日本バレーボール殿堂入りを果たしました。

Wikipediaより

大松 博文 (だいまつ ひろふみ^[2]、1921年2月12日 - 1978年11月24日) は、日本のバレーボール指導者、政治家、参議院議員 (1期)。香川県綾歌郡宇多津町出身。

ニチポ-貝塚監督時代は、自ら育て上げた東洋の魔女を率いて、辞任までに175連勝を記録。1962年 (昭和37年)、第4回女子世界選手権でソ連を破り優勝。国民的英雄として迎えられる。1964年東京オリンピックでは全日本女子を監督として金メダルに導く^[3]。「回転レシーブ」などを考案、その過酷な練習で「鬼の大松」ともいわれた。情報誌『imidas2001』 (集英社) の「20世紀を創った人々550」では、バレーボールの分野で前田豊、猫田勝敏と並んで3人のうちの1人に数えられた。

Google検索結果

"石井嘉幸" バレーボールとの一致はありません。

石井嘉幸 バレーボール の検索結果 (引用符なし):

検索条件と十分に一致する結果が見つかりません。

探しているページに表示されている可能性がある言葉で検索してみてください。たとえば、「ケーキの作り方」ではなく、「ケーキのレシピ」でお試しください。

お困りの場合は、Google での検索に関するその他のヒントをご覧ください。

そもそも実在しないバレーボールプレーヤー

[おまけ] 誤情報と創作 は表裏一体

- 「偽情報 (Hallucination)」の要因と「新アイディアの提示/物語生成」といった創作能力の根源は同じ
 - => 用途の違いで「良い」と「悪い」の判断が反転

偽情報 (Hallucination)

1 パレーボールで回転レシーブを考案したのは誰ですか?

2 パレーボールで回転レシーブを考案したのは、日本の元パレーボール選手である石井嘉幸 (いしいよしゆき) 氏です。石井嘉幸は、1990年代に日本の男子パレーボール代表として活躍し、特にその優れたレシーブ技術で知られていました。彼は独自のスタイルである回転レシーブを開発し、これは後に多くのプレーヤーやチームに影響を与えました。その後、パレーボール界での彼の功績が認められ、日本パレーボール殿堂入りを果たしました。

Wikipediaより

大塚 博文 (だいまつひろみ) 1921年2月12日 - 1978年11月24日) は、日本のパレーボール指導者、政治家、朝日新聞記者 (1期)。香川県高松市多度町出身。

ニデボ一級球監時代は、自ら育て上げた藤野の腕を率いて、計年までに175連勝を記録。1962年 (昭和37年)、第4回女子世界選手権でソ連を破り優勝、国民的英雄として迎えられる。1964年東京オリンピックでは全日本女子を監督として金メダルに導く。『回転レシーブ』などを考案。その逸話な経歴で『鬼の大塚』ともいわれた。情報誌『Fridays2001』 (集英社) の「20世紀を創った人々550」では、パレーボールの分野で前田勲、藤田勝敏と並んで3人のうちの1人に数えられた。

Google検索結果

"石井嘉幸"パレーボールとの一致はありません。

石井嘉幸 パレーボール の検索結果 (引用符なし):

検索条件と十分に一致する結果が見つかりません。

表示しているページに表示されている可能性がある言葉で検索してみてください。たとえば、「ケーキの作り方」ではなく、「ケーキのレシピ」でお試ください。

お探しの場合は、Google での検索に関する他のヒントをご覧ください。

そもそも実在しないパレーボールプレーヤー

創作 / 新アイディア

1 パレーボールの回転レシーブを題材に創作して (200字以内)

2 タイトル: 「舞台の上で踊り出す回転レシーブ」

舞台は美しく輝くスポットライトに照らされ、公のユウリは、無名の高校パレーボールチームでいます。

試合が始まり、強力なサーブが飛び交います。足取りでボールを追いかけます。しかし、相手そんな中、ユウリはふと一瞬、試合以外の想トレーニングや笑い声が、彼女の心を変えて

彼女は決意を新たに、瞬間移動のようにボールを動かして、次々と敵の攻撃を跳ね返します。り動きです。

ユウリの回転レシーブは、仲間との絆と努力、試合を盛り上げる存在となったのです。に、観客たちは感動し、応援に熱が入ります

そしてついに迎えた最後の一球、ユウリの回って、彼女の背後から誰かが力強くスライク回転レシーブは、新たな伝説として語り継がれることでしょう。

1 今までにない新しいお菓子を新発案したいのだけれど、アイデアが浮かばないです。今までにない食べ方にも似た目のインパクトがあるお菓子のアイデアのブレインストーミングをしたいのですが、まずたまたま自然になるアイデアをだしてもらえないですか?

もちろんです! 以下に、見た目のインパクトがある新しいお菓子のアイデアのいくつかをご提案します。これらはブレインストーミングのたたき台としてご利用いただければ幸いです。

- カラフルな宇宙ケーキ: レインボーカラーのスポンジケーキやクリームを使い、宇宙のような美しい模様を作り出します。食べるときに宇宙をイメージできるような、幻想的なデザートです。
- ジュウリウム チョコレートで作られた小さなデザート (種物の糖) を作り、内部には様々なフレーバーのチョコレートやカラフルなキャンディが詰まっており、食べるたびにデザートが壊れて中のお菓子が現れる仕掛けになっています。
- 立体道路クッキー: 立体的な道路の形をしたクッキーを作り、クッキーの中には果物やチョコチップが隠れており、食べながら道路を歩いていく楽しみがあります。
- ジュエルデザート: キラキラと輝くようなデザートを作り、透明なグレースや糖でコーティングし、表面にはカラフルな食用ジュエルを散りばめます。食べるたびにまるで宝石を食べているような感覚が味わえます。
- ピクセルクッキー: クッキーの表面に小さなピクセルアートを作り、各ピクセルには違う味や色のフロスティングが使われており、食べるたびに違う組み合わせの味が楽しめます。

これらのアイデアは、見た目のインパクトがあり、食べる楽しみや驚きを提供することを目的としています。お好みやアレルギーに合わせたアイデアに合わせてカスタマイズしていただければ幸いです。どのアイデアが気に入りましたか?

日本語のことがわかっているか?

- 学習データの大半が英語データ
 - GPT-3の日本語データはわずか0.1%
- 日本語の単語区切りはわかっていない
 - <https://platform.openai.com/tokenizer>

	language	number of words	percentage
1	English (en)	181,014,683,608	92.65%
2	French (fr)	3,553,061,536	1.82%
3	Germany (de)	2,870,869,396	1.47%
4	Spanish (es)	1,510,070,974	0.77%
5	Italiano (it)	1,187,784,217	0.61%
6	Portuguese (pt)	1,025,413,869	0.52%
7	Dutch (nl)	669,055,061	0.34%
8	Russian (ru)	368,157,074	0.19%
9	Romanian (ro)	308,182,352	0.16%
10	Polish (pl)	303,812,362	0.16%
11	Finnish (fi)	221,644,679	0.11%
12	Danish (da)	221,551,540	0.11%
13	Swedish (sv)	220,920,577	0.11%
14	Japanese (ja)	217,047,918	0.11%
15	Norwegian (no)	212,193,299	0.11%
16	Chinese (zh)	193,517,396	0.10%
17	Czech (cs)	139,918,438	0.07%
18	Hungarian (hu)	127,224,375	0.07%

日本の首都は東京です。

岸田総理は11時ごろに首相官邸を出発する見込みです。

今日も桜は満開です。神奈川県は冷え込みが激しいです。

日本語 (処理単位が文字またはバイトレベル)

日本の首都は00京です。

0000000理は11時ごろに首相00000を出00する見00みです。

今日も00は00開です。0000000県は00え00みが00しいです。

英語 (処理単位がほぼ単語)

Many words map to one token, but some don't: Indivisible. Unicode characters like emojis may be split into many tokens containing the underlying bytes: 🍌🍌🍌 Sequences of characters commonly found next to each other may be grouped together: 1234567890

文章生成AIが苦手なこと

- 新情報や情報の追従
 - 例：ChatGPT => 2021年までのデータで事前学習
 - GPT-4は 2023.3までのデータで学習
 - 最新の情報は学習されていない可能性が高い
- 地域に限定した情報や特定のコミュニティにしか通用しない話題など
 - 一般的な知識を獲得している
- 唯一の回答を答える
 - => 一般論的かつ複数の回答があるような質問には的確な回答ができる場合が多い

まとめ：文章生成AIの特性

- 詳細情報は非公開
- 非常に高い到達点
 - => 多くの新たな問題も浮き彫りにした
- 誤情報の問題
 - 文章生成AIの特性に起因
- 苦手なことも多くある
 - 日本語（英語以外の言語）
 - 新しい情報，変化する情報

[参考] 日本語に強い言語モデルの開発

- LLM-jp : 国立情報学研究所 (NII) が主導
- All Japanで産学の自然言語処理と関連分野の800名以上規模の研究者/技術者が参加 (産学半々程度)
- オープンかつ日本語に強い大規模モデルを構築し、LLMの原理解明に取り組む
- モデル・データ・ツール・技術資料等の成果物を議論の過程・失敗を含め**すべて公開**

<https://llm-jp.nii.ac.jp>



- 導入 : 文章生成AI とは ?
- 文章生成AI の成り立ち
- 文章生成AI の特性
- 文章生成AI の展望
- まとめ

文章生成AIによる社会変革

- 産業「革命」級の出来事?

➔ 大きな社会変革を促す?

日報/報告書
企画書
説明書
コールセンター/
ヘルプデスク/
クレーム対応

言語を使わない産業はない / 人間は思考する時に言語を用いる

- 不可逆かつ革命的な社会変革の受け入れ
 - 文章生成AIありきの世界を模索する必要性

これまでも自動車 / インターネット / スマホなどが登場するたびに人間はそれらがあることが当たり前の生活に自然と適合してきた

[参考] 生成系AI：画像生成との対比

- 画像生成AI
 - 比較的適応分野が限定
 - 目的に特化
- 文章生成AI
 - 広範/雑多な目的
 - 日常使い

より専門家



秘書/執事
ジェネラリスト



今後想定される社会的課題

- **不適切利用** (技術的欠陥の利用など)
 - 意見/印象の操作
 - 生成系AIのハッキング手法の公開や共有
 - 意図的な誤った情報による汚染
- **データ汚染 (Contamination)**
 - 文章生成AIの誤生成がweb上に溢れかえると誤情報が真実として信じられる可能性
- **著作権侵害**
 - 意図せず他人の権利を侵害する可能性
- **更なる情報格差**
 - 生成AIが使いこなせるかで更なる格差を生む可能性
- **環境問題 (エネルギー)**
 - ChatGPTなどの構築/利用には膨大な電力が必要
 - 持続可能社会のための取り組みが必須
- **生成系AIネイティブへの教育や社会的価値の在り方**
 - 生成系AIのサポートを得られる時代における「価値」とは？
- **利用に関する法整備**
 - 法整備には協議が必要 / 時間がかかる
 - 技術発展の速さに追従困難

汎用人工知能

- 人工知能研究者の究極のゴール の一つ

(対話型)文章生成AI => 汎用人工知能??

ChatGPTがもたらした/もたらすもの

- 汎用人工知能への**第一歩**を示した (歴史を作った)

AIの区分	
特化型人工知能 (Narrow AI) => これまでのほぼ全てのAI	汎用人工知能 (Artificial General Intelligence: AGI) => 人間と同等の知的能力

- 汎用人工知能の実現に向けて**さらなる深化**を促す
 - 現在の自然言語処理研究の多くはこの方向性を旨す

[参考] AGIのレベル定義

<https://arxiv.org/abs/2311.02462>

Performance (rows) x Generality (columns)	Narrow	General
	clearly scoped task or set of tasks	wide range of non-physical tasks, including metacognitive abilities like learning new skills
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging equal to or somewhat better than an unskilled human	Emerging Narrow AI GOFAl4; simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	Emerging AGI ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023)
Level 2: Competent at least 50th percentile of skilled adults	Competent Narrow AI toxicity detectors: Jigsaw (Das et al., 2022); Smart Speakers: Siri, Alexa, Google Assistant; VQA systems: PaLI (Chen et al., 2023), Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert at least 90th percentile of skilled adults	Expert Narrow AI Grammarly (Grammarly, 2023); Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	Expert AGI not yet achieved
Level 4: Virtuoso at least 99th percentile of skilled adults	Virtuoso Narrow AI Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016, 2017)	Virtuoso AGI not yet achieved
Level 5: Superhuman outperforms 100% of humans	Superhuman Narrow AI AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	Artificial Superintelligence (ASI) not yet achieved

ChatGPTがもたらした/もたらすもの

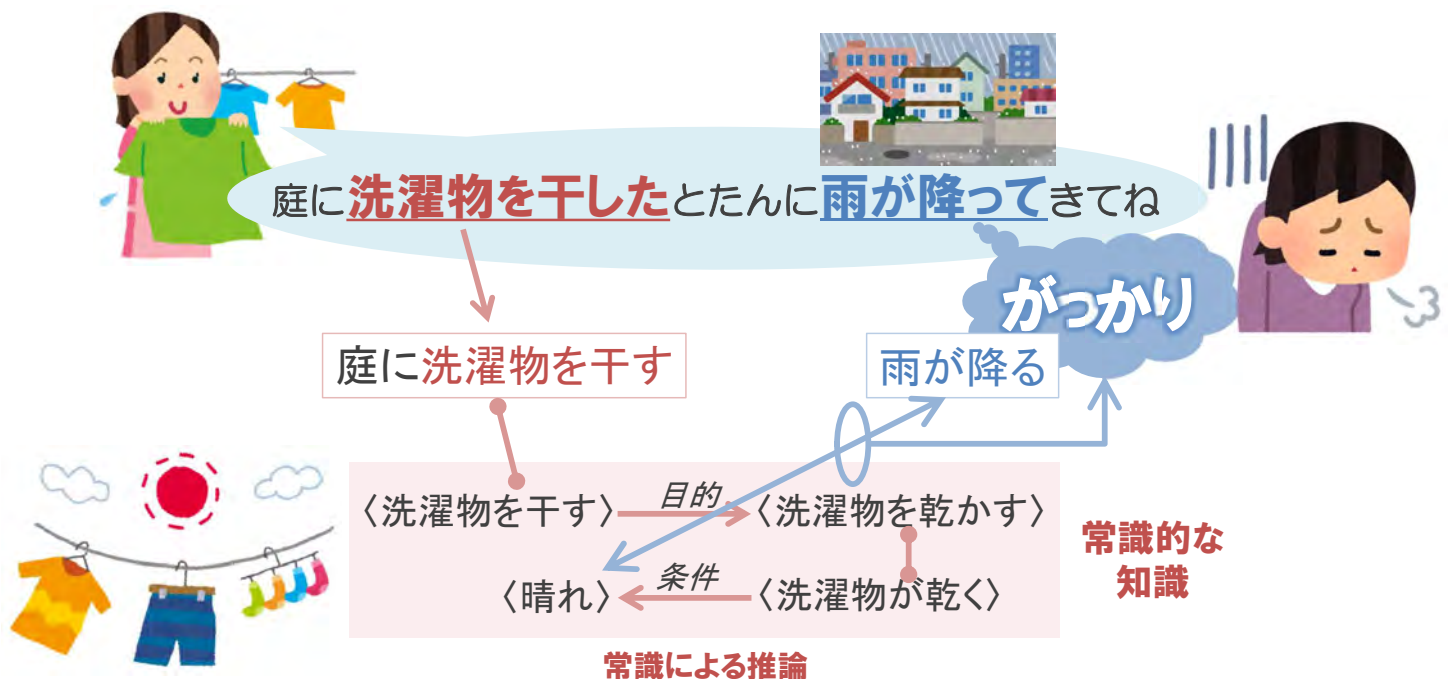
(著名人でも様々な意見はありますが)

現時点においては

- 人間と同じ「知性」を持った何か **ではない**
- シンギュラリティにつながる何か **ではない**
- 人間の本質的な役割を奪う何か **ではない**
- あくまで人間にとって負担の大きい労働などを代替したりサポートすることが役割 **というスタンス**

言葉を理解するとは？

言語の真の理解 => 文脈 / 状況の理解も含む
+ 人間が持つ常識 / 感情 も正しく扱う必要



知能? 知性?

- 文章生成AIの生成した文章に**知性**を感じるか?
Yes / No
(著名人の中でも多くの意見がある質問)
- **Theory of Mind / 心の理論**
 - 「他者が感じ考えていること」を推測
 - 最近 機械学習や自然言語処理研究で注目されるように
 - 問い：文章生成AIは Theory of Mind をもつか??

本講演では議論しない

まとめ：文章生成AIの展望

- **産業革命級**の出来事 => **社会変容**を促す
 - 人間の知的活動には言葉が関わる (関わらないことはない)
- **汎用人工知能 (AGI)**
 - 汎用人工知能：AGIの入り口に立った?
 - ただし真の汎用知能に向けては問題が山積み状態
 - 現状は人間と同じ理解/手続きはとっていない(ようだ)
- **言葉を真に理解するとは?**
 - 文脈 / 状況 / 感情 なども含めた正しく理解が必要
 - **Theory of Mind / 心の理論**
 - 昨今の対話型文章生成AIは知性があると言えるのか?
=> ここでは議論しない

- 導入：文章生成AIとは？
- 文章生成AIの成り立ち
- 文章生成AIの特性
- 文章生成AIの展望
- まとめ

まとめ

まとめ：文章生成AIとは？

- 文章生成AIを一言で説明すると？
 - (人が人に話すような) 対話形式の指示を受け付け
その指示に適した文章を生成する文章生成器
- 人間の様々な知的活動をサポート可能
=> 使い方の工夫次第で無限の可能性!?
- ChatGPT

まとめ：文章生成AIの成り立ち

- 土台：言語モデル
 - 文章の出現確率を予測する確率モデル
 - ニューラル言語モデル：文章の出現確率を深層ニューラルネットワークで近似
- 言語モデルの事前学習と微調整学習により実現
 - 事前学習：記憶
 - 微調整学習：記憶の呼出

文章生成AI (ChatGPT系ツール) と
うまく付き合い
賢く使いこなしていきましょう

まとめ：文

- 詳細情報
- 非常に高い到達点
 - => 多くの新たな問題も浮き彫りにした
- 誤情報の問題
 - 文章生成AIの特性に起因
- 苦手なことも多くある
 - 日本語 (英語以外の言語)
 - 新しい情報, 変化する情報

● 汎用人工知能 (AGI)

- 汎用人工知能：AGIの入り口に立った？
 - ただし真の汎用知能に向けては問題が山積み状態
 - 現状は人間と同じ理解/手続きはとっていない (ようだ)
- 言葉を真に理解するとは？
 - 文脈 / 状況 / 感情 なども含めた正しく理解が必要
 - Theory of Mind / 心の理論
 - 昨今の対話型文章生成AIは知性があると言えるのか？
=> ここでは議論しない