LLM-jp

黒橋 禎夫

国立情報学研究所 所長/京都大学 特定教授

言語処理学会理事会主催 緊急パネル

緊急パネル:ChatGPTで自然言語処理は終わるのか?

- ファシリテーター
 - 乾健太郎氏(東北大)
- ・パネリスト
 - 黒橋 禎夫氏(京大)
 - o 相良 美織 氏(バオバブ)
 - 佐藤 敏紀 氏(LINE)
 - 鈴木潤氏 (東北大)
- 2023年 谷中瞳氏(東大)



- 3月14(火) 13:10-13:50, H会場(劇場ホール)
 - Slack: #0314-1310-緊急パネル-h会場(質問はこちらへ)
 - Twitter: #ChatGPTで自然言語処理は終わるのか

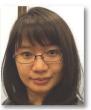












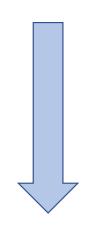
ChatGPTで自然言語処理は終わるのか?

黒橋 禎夫

京都大学教授/NII 所長特別補佐 (2023/4から NII 所長/京都大学特定教授)

言語処理学会第29回年次大会 緊急パネル (2023/03/14)

ChatGPTで自然言語処理は終わるのか?

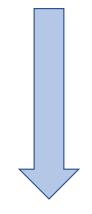






旧来のNLP

形態素解析、構文解析、省略解析 ... 関係抽出、要約、MT、対話システム ...



数年後には、1年後にも

新しい自然言語処理

もっと人間や社会へ向かうもの

- 脳神経科学との関係
- 人の感情
- 人の理解、納得、説得
- Al Native世代の教育、批判的思考の涵養
- 情報の信頼性、関係付け(参照)

そのためには

- LLMが完全にオープンであること
- また、ChatGPTにおける日本語コーパスは圧倒的に少ない
- 日本文化や経済安全保障の視点
- ・良質で十分な日本語テキスト (+英語テキスト) で学習したオープンな大規模言語モデルを日本が 持つ必要性

まとめ

• ChatGPTをディスっても仕方がない

旧来のNLPは終わるが、新しいNLPが始まる (ただしそれをNLPとよぶかは?)

みんなでオープンなLLMを作ろう! (我が国でも大きなムーブメントとなりつつある)

LLM-jp (LLM勉強会)

- 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・ 技術資料等を議論の過程・ 失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可

2023.5

自然言語処理の研究者の有志約30名による勉強会を開催

LLM研究開発はビッグサイエンス

コーパス構築WG

大規模コーパス



河原大輔教授 (早稲田大学)

安全性WG



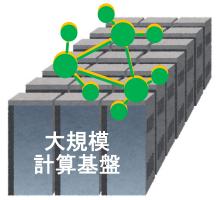
関根聡特任教授(NII)

マルチモーダル WG



岡崎直観(東工大)

モデル構築WG

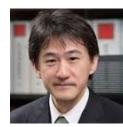




横田理央教授 (東工大)



鈴木潤教授 (東北大)



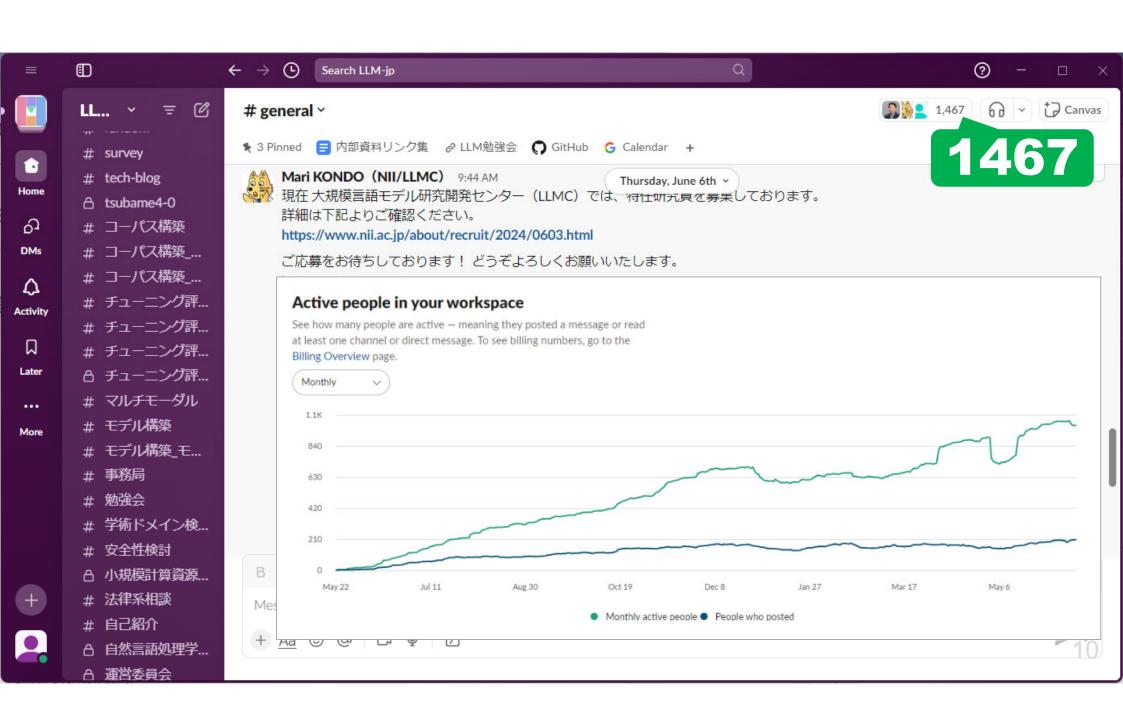
田浦健次郎教授 (東大)

チューニング評価WG





宮尾祐介教授(東大)



LLM-jp (LLM勉強会)

- 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・ 技術資料等を議論の過程・ 失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可

2023.5

自然言語処理の研究者の有志約30名による勉強会を開催

2023.10

mdxを用いて130億パラメータの大規模 言語モデル「LLM-jp-13B」を公開

• mdx: データ活用社会創成プラットフォーム. 9大学2研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境

LLM-jp モデル

v1.0 (2023/10/20 公開) ※ githubのダウンロード件数 3万超

- アーキテクチャ: GPT/トークナイザ: 語彙数50k
- 事前学習コーパス:日本語,英語,コードを混ぜた約300Bトークン
- チューニング用データセット:jaster, dolly (日), open assistant (日). 185k サンプル

v1.1 (2024/2/9 公開)

- v1.0 のチューニング手法を改良
 - Ichikara instruction データ, 3k サンプルを追加(jasterは不使用)
 - Learning from Human Feedback 導入, 12k サンプル

v2.0 (2024/4/30 公開)

- アーキテクチャ:Llama/トークナイザ:語彙数100k
- 事前学習コーパス:日本語WebデータをmC4 に代えてCommon Crawl をクリーニングして使用
- チューニング用データセット:dolly (日英), open assistant (日英), Ichikara(日), 75k サンプル. さらに安全性についてAnswerCarefully データを構築

根拠を調べるためのコーパス検索

@llmjp-13b-finetuned 地震の原因はナマズが動くことですか?

1 reply



Ilmip-13b-finetuned APP 3 hours ago

いいえ、地震はナマズのせいではありません。地震は、地殻に蓄積されたエネルギーが突然 解放されることで起こる自然現象です。地下の岩盤が、圧力に耐えきれずにずれ始めると、 断層に力がかかり、地震が発生する。

地震は通常、海底の隆起や火山の噴火など、特定の要因が引き金となって起こる。地震学 [...]

#1

訓練 コーパス (260B)

Dataset: Wikipedia

URL: https://ja.wikipedia.org/wiki/地震

Training step: 28,955 (/61,000)

地震の原因論とメカニズム論の展開神話など日本 (大和民族) では古来より「地中深くに大ナマズが 存在し、その大ナマズが暴れることにより大地震が 起きる」という俗説が信じられていた。現代におい てもよく知られた俗説だが、ナマズが地震を予知で きる根拠は見つかっていない[...]

Dataset: Common Crawl

#2

URL: https://gbank.gsj.jp/wellweb/GSJ/kaisetsu/science/science.html

Training step: 168 (/61,000)

日本は世界有数の地震国です。そこで誰でも考える のは、「地震発生が事前にわかればあれほどの被害 は出ないのに」ということ。つまり地震予知です。 この地震予知に関しては昔からいろいろな事が言わ れてきました。例えば「地震の前にはナマズが暴れ る」、「地震の前には[...]

土佐日記の現代語への翻訳

ChatGPT 4o

以下の文を現代語に変換して。

ある人、県の四年五年果てて、例のことどもみなし終へて、解由など取りて、住む 館より出でて、船に乗るべき所へわたる。かれこれ、知る知らぬ、送りす。

⑤ もちろんです。以下の文を現代語に変換いたします。

ある人が、県での任期を四、五年務め終え、通常の手続きをすべて済ませて、解由状などを受け取り、住んでいた館を出て、船に乗る場所へ向かう。あちらこちらで、知っている人も知らない人も見送りに来る。

LLM-jp 13B v2



Sadao Kurohashi (NII/京大) 6 minutes ago

@llmjp-13b-finetuned 以下の文を現代語に変換してください。ある人、県の四年五年果てて、例のことどもみなし終へて、 解由など取りて、住む館より出でて、船に乗るべき所へわたる。かれこれ、知る知らぬ、送りす。

1 reply



Ilmjp-13b-finetuned APP 6 minutes ago

ある人が、県の四年と五年を終えて、これまでのことをすべて片付けて、解由(かりゆ)などを受け取って、住んでいる館から出て、船に乗るところへ行くところへ行く。みんなが、知った知らないと言いながら見送る

LLM-jp (LLM勉強会)

- 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・ 技術資料等を議論の過程・ 失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可

2023.5

自然言語処理の研究者の有志約30名による勉強会を開催

2023.10

mdxを用いて130億パラメータの大規模 言語モデル「LLM-jp-13B」を公開

2023.11

ABCI第2回LLM構築支援プログラム採択 1750億パラメタモデルの学習に着手

2024.1

GENIAC 第1期採択(2024.4から 1750億パラメタモデルを学習)

- mdx: データ活用社会創成プラットフォーム、9大学2研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境
- ABCI: AI橋渡しクラウド. 産業技術総合研究所(AIST)が提供するAI向け計算用で現状国内最大の計算資源
- GENIAC: Generative AI Accelerator Challenge.日本国内の基盤モデル開発力の底上げのために計算資源の提供等を行う経産省のプログラム

文部科学省:生成AIモデルの透明性・信頼性の確保に向けた

研究開発拠点形成(事業期間:令和5年度~令和10年度)

大規模言語モデル研究開発センター(LLMC)

LLM-jpの活動

- LLMC主宰の活動と位置づけ
- 1400名超の参加者
- 完全にオープン

LLMCの活動

- LLMの構築・公開
- LLMの透明性・信頼性確保に関する 研究開発
- LLMの高度化に関する研究開発
 (データ開拓・開発、企業との共同研究等では一部クローズな活動もありえる)



If you want to go fast, go alone If you want to go far, go together

