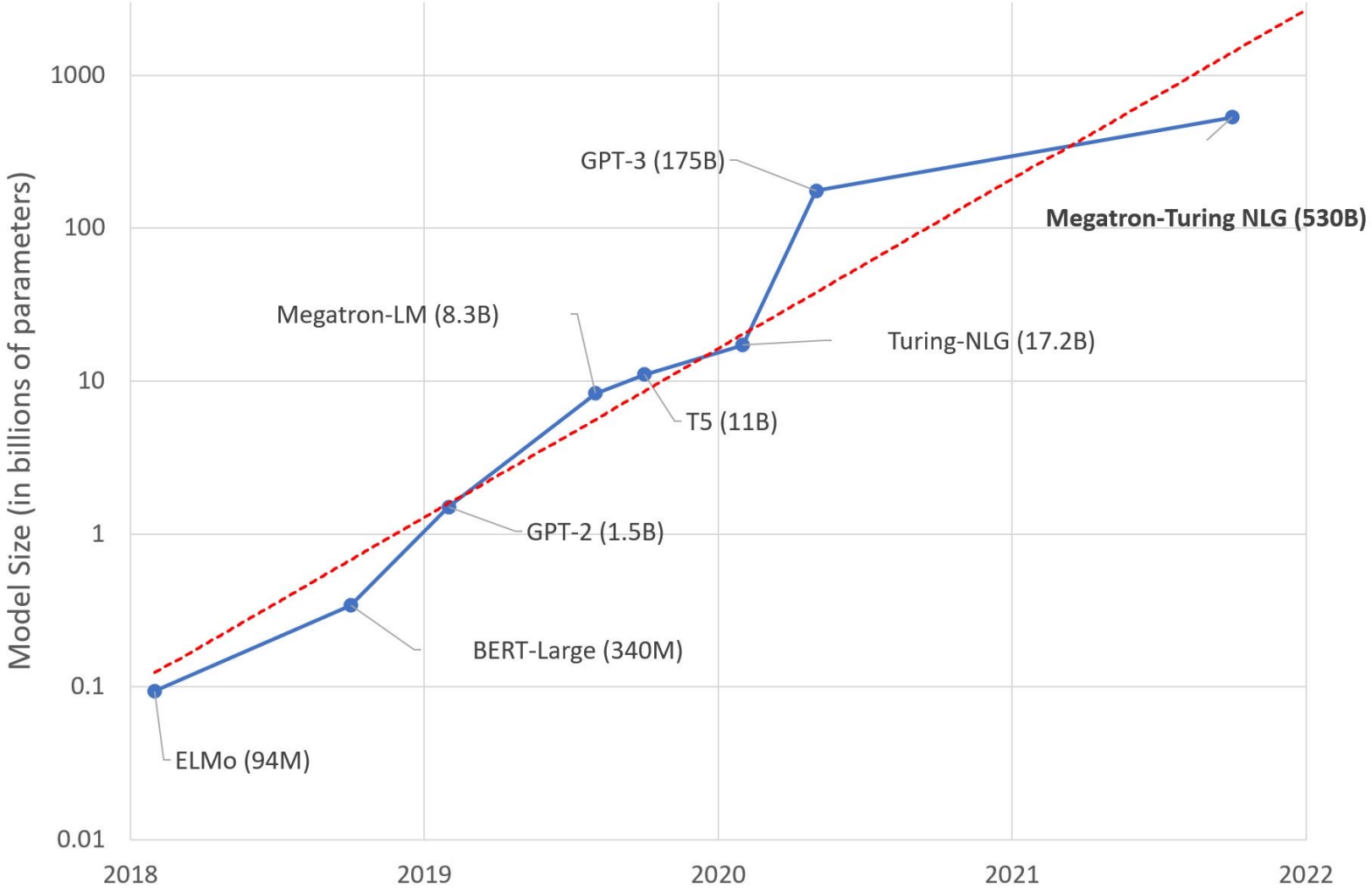
日本語BigBirdの構築

河原大輔
早稲田大学

ABCIグランドチャレンジ2022年度第3回参加メンバー:
近藤瑞希, 王昊, 井手竜也, 伊藤俊太朗, Ritvik Choudhary, 栗原健太郎, 河原大輔

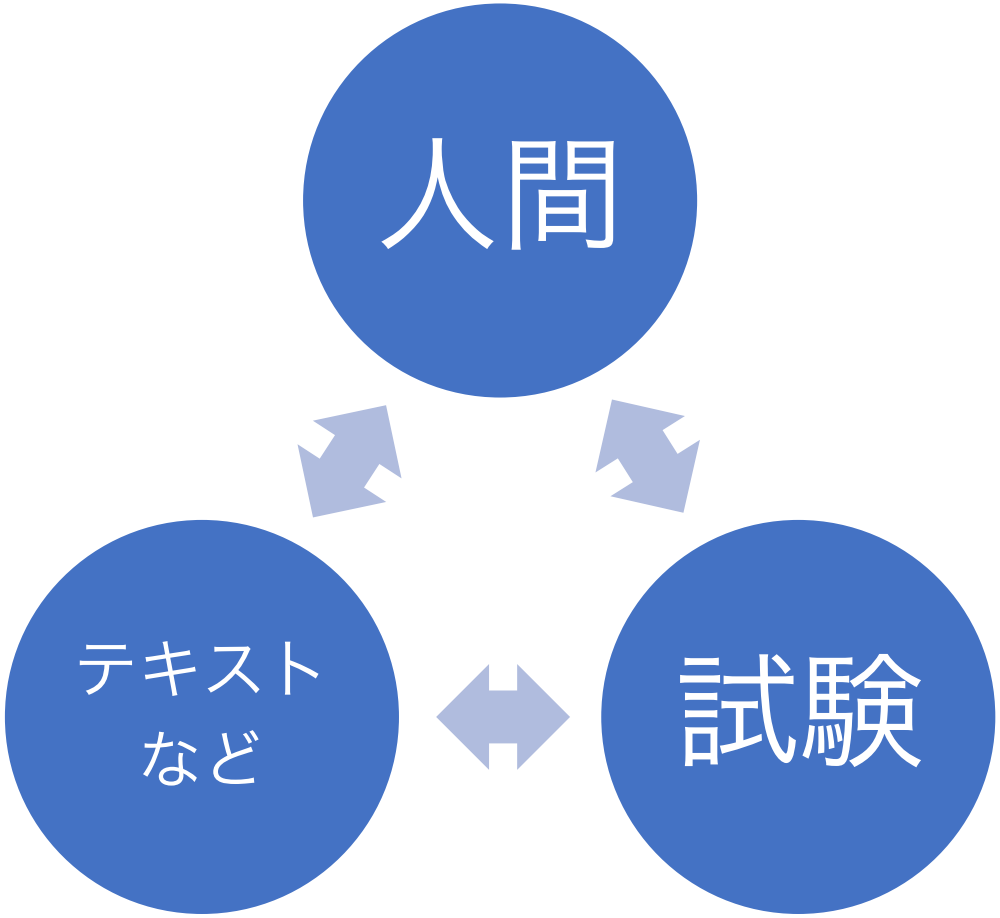
ABCIグランドチャレンジ成果報告会 (2024/1/11)

大規模言語モデル(LLM)の進展

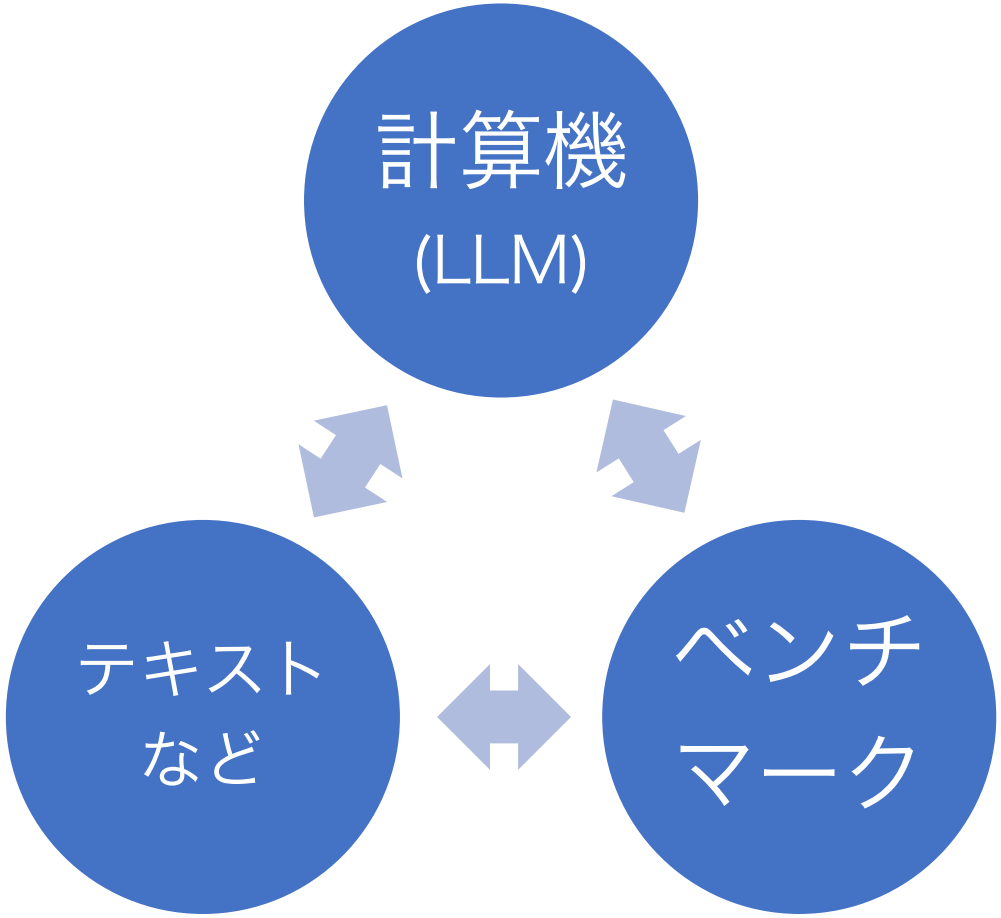


https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

人間の勉強過程



計算機の学習過程



日本語言語理解ベンチマークJGLUE (v1)

[栗原+ 2022] [栗原+ 2023]

- GLUEやSuperGLUEのタスクを広くカバーするように構成
- 構築にはYahoo!クラウドソーシングを利用

タスク	データセット	train	dev	test
文章分類	MARC-ja	187,528	5,654	5,639
	JCoLA [染谷+ 2022]	-	-	-
文ペア分類	JSTS	12,451	1,457	1,589
	JNLI	20,073	2,434	2,508
QA	JSQuAD	62,859	4,442	4,420
	JCommonsenseQA	8,939	1,119	1,118

ヤフー・早大共同研究　https://github.com/yahoojapan/JGLUE

JGLUEデータ例 (1/2)

MARC-ja

positive

色も履き心地も最高です。私の場合は夏場の船釣りに使っています。

negative

何このSDカードデーター移せないしコマンドで調べたらエラー出るしこれはない

positive → negative

単純に一週間の天気を知りたいのであればこれで十分。だがこの程度で有料はいかがなものか

クラウドソーシングでの回答
positive: 0, negative: 10

JSTS/JNLI

類似度: 4.4, 推論関係: entailment

文1: 街中の道路を大きなバスが走っています。
文2: 道路を大きなバスが走っています。

類似度: 3.0, 推論関係: neutral

文1: テーブルに料理がならべられています。
文2: テーブルに食べかけの料理があります。

類似度: 2.0, 推論関係: contradiction

文1: 野球選手がバットをスイングしています。
文2: 野球選手がキャッチボールをしています。

JGLUEデータ例 (2/2)

JSQuAD

[タイトル] 東海道新幹線
1987年（昭和62年）4月1日の国鉄分割民営化により、JR東海が運営を継承した。西日本旅客鉄道（JR西日本）が継承した山陽新幹線とは相互乗り入れが行われており、東海道新幹線区間のみで運転される列車にもJR西日本所有の車両が使用されることがある。2020年（令和2年）3月現在、東京駅 - 新大阪駅間の所要時間は最速2時間21分最高速度285km/hで運行されている。

質問: 2020年、東京～新大阪間の最速の所要時間は
答え: 2時間21分

質問: 東海道新幹線と相互乗り入れがされている路線はどこか？
答え: 山陽新幹線

JCommonsenseQA

問題:　会社の最高責任者を何というか？
選択肢: 教師, 部長, 社長, 部下, バイト

問題:　スープを飲む時に使う道具は？
選択肢: スプーン, メニュー, 皿, フォーク, はし

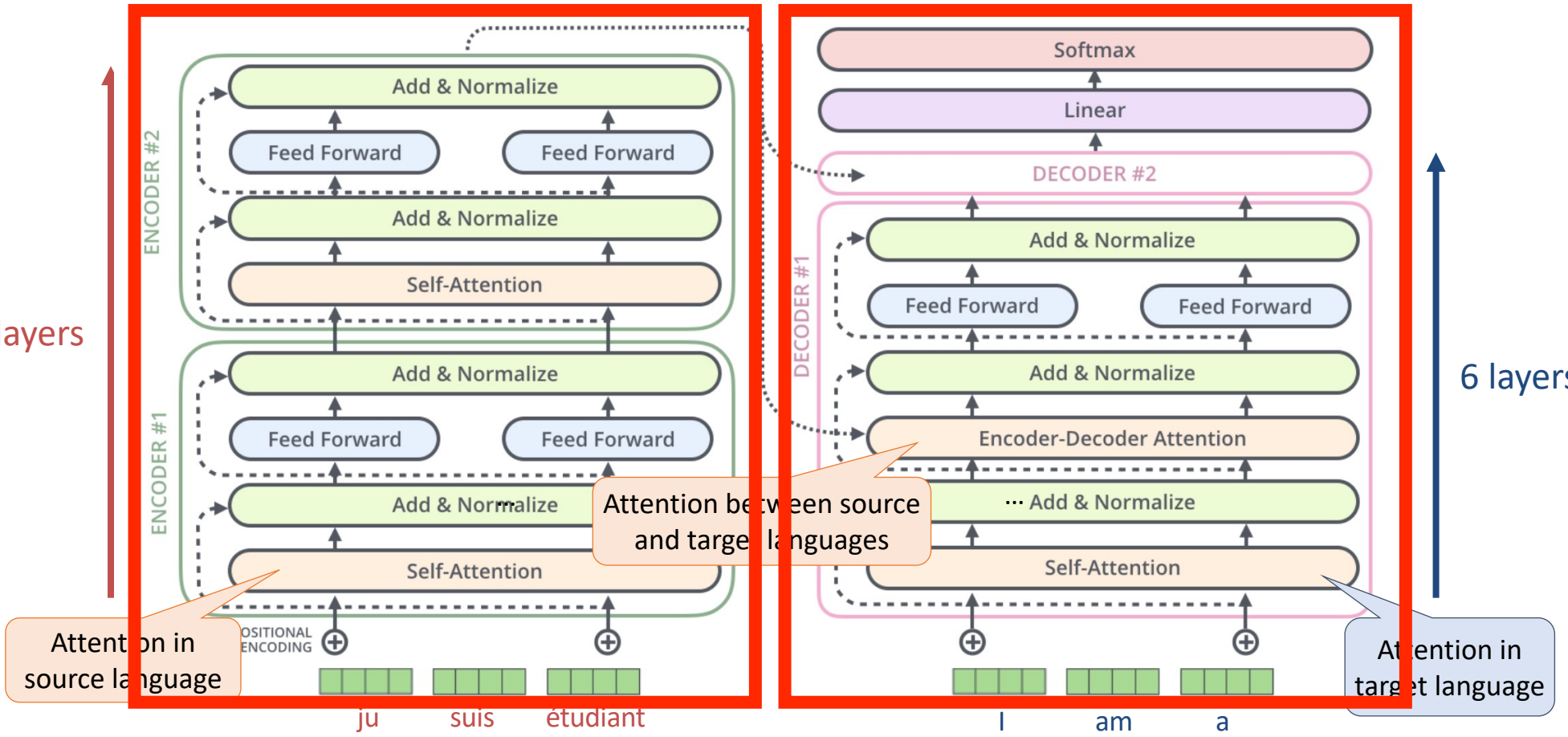
https://jalammar.github.io/illustrated-transformer/

LLMの分類

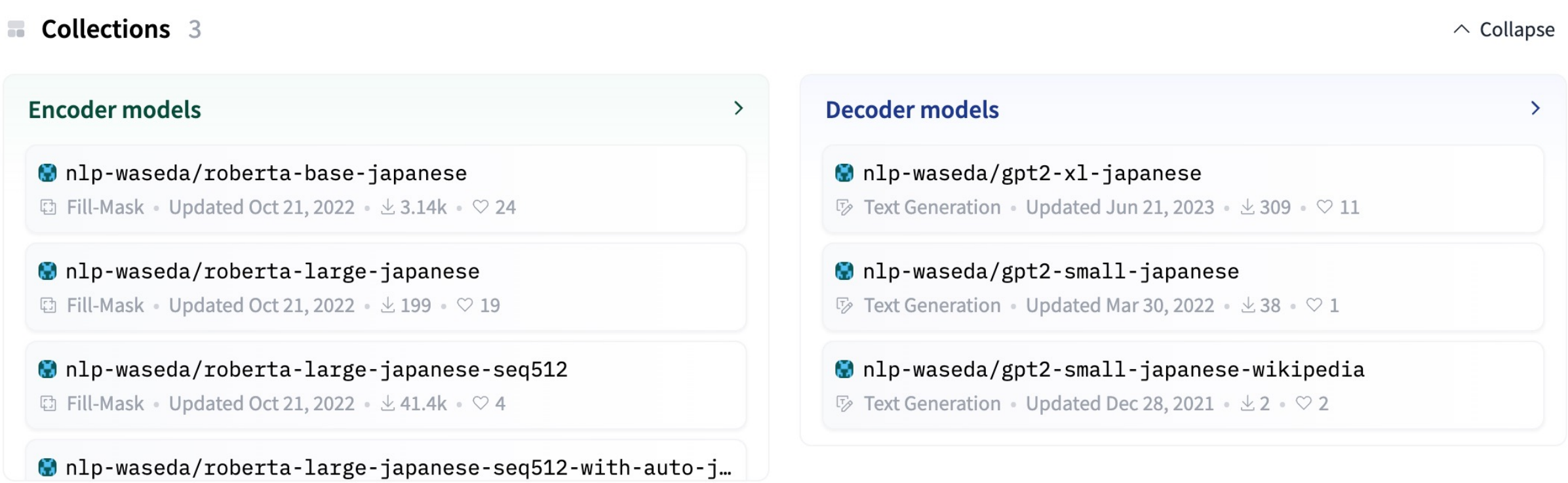
エンコーダ・デコーダ (T5など)

エンコーダ (BERT系)

デコーダ (GPT系)



nlp-wasedaモデル群



https://huggingface.co/nlp-waseda

nlp-waseda/roberta-{base, large}-japanese

- モデル
  - base: 110Mパラメータ
  - large: 330Mパラメータ
- 事前学習テキスト (約10Bトークン)
  - 日本語Wikipedia + CC-100日本語部分
- 計算資源
  - ABCI Aノード x 1 (A100 x 8)
    - base: 1週間 (500ポイント)
    - large: 2週間 (1,000ポイント)

※ 試行錯誤が必要で、実際には
3倍程度のポイントを使用

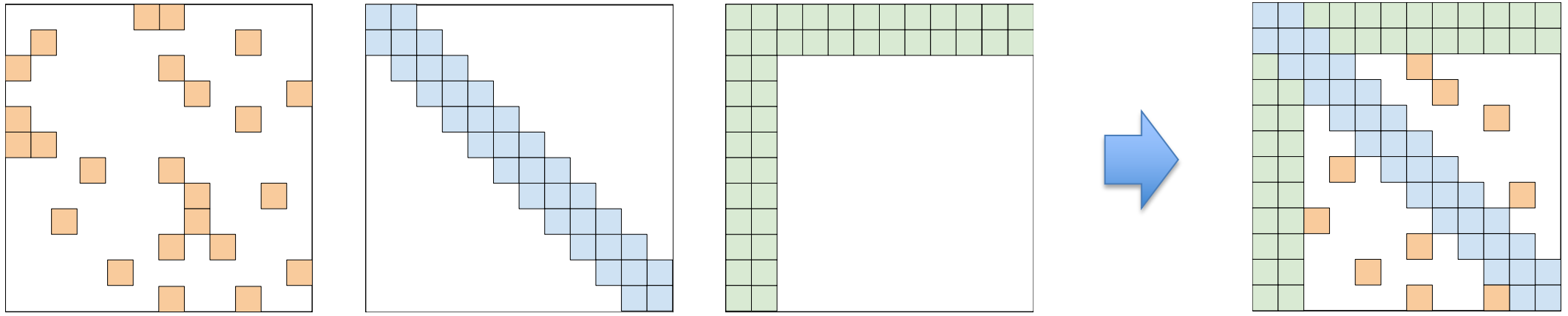
日本語を入力できるLLMの最大コンテキスト長

- エンコーダ
  - BERT: 512トークン
  - RoBERTa: 512トークン
  - LUKE: 512トークン
  - DeBERTa: 512トークン
  - BigBird: 4,096トークン
- デコーダ
  - GPT: 512トークン
  - GPT-2: 1,024トークン
  - GPT-3: 2,048トークン
  - GPT-3.5: 4,096トークン
  - GPT-4: 8,192トークン

長文の文章読解や抽出型要約に必要

ABCIグランドチャレンジ2022 (1/2)

- 日本語BigBirdの学習
  - 長い系列(4,096トークン)の入力に対応したモデル [Zaheer+ 2020]



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

ABCIグランドチャレンジ2022 (1/2)

- 日本語BigBirdの学習
  - 長い系列(4,096トークン)の入力に対応したモデル [Zaheer+ 2020]
  - baseサイズ (110Mパラメータ)
- 事前学習テキスト (約48Bトークン)
  - 日本語Wikipedia + CC-100日本語部分 + OSCAR日本語部分
- 計算資源
  - ABCI Aノード x 120 (A100 x 960), 24時間

ABCIグランドチャレンジ2022 (2/2)

- 結果: 失敗
  - リハーサルでは起きなかった様々なエラーが発生
    - ファイルディスクリプタ不足
    - DeepSpeed CPUオフロードのエラー
  - 想定よりスケールしなかった
    - Hugging Face + DeepSpeedではスケールしない?
- その後
  - ABCI Aノード x 2 (A100 x 16), 2週間で学習を完了

日本語(大規模)言語モデルとその性能

	MARC-ja (acc)	JSTS (Pearson)	JNLI (acc)	JSQuAD (F1)	JComQA (acc)
(人間)	0.989	0.899	0.925	0.944	0.986
東北大 BERTBASE	0.958	0.909	0.899	0.941	0.808
東北大 BERTBASE (文字)	0.956	0.893	0.892	0.937	0.718
NICT BERTBASE	0.958	0.910	0.902	0.947	0.823
早稲田大 RoBERTaBASE	0.962	0.913	0.895	0.927	0.840
▶ 早稲田大 BigBirdBASE	0.959	0.888	0.896	0.933	0.787
Studio Ousia LUKEBASE*	0.965	0.916	0.912	-	0.842
京大 DeBERTaBASE*	**0.970**	**0.922**	**0.922**	**0.951**	**0.873**
東北大 BERTLARGE	0.955	0.913	0.900	0.946	0.816
早稲田大 RoBERTaLARGE (s512)	0.961	0.926	0.926	**0.963**	0.891
Studio Ousia LUKELARGE*	0.965	**0.932**	**0.927**	-	**0.893**
京大 DeBERTaLARGE*	**0.968**	0.925	0.924	0.959	0.890

* http://huggingface.co/ にある各モデルの記述より

言語理解ベンチマークの改良に向けて

- より難易度の高いベンチマークの構築: JGLUE v2
  - cf. SuperGLUE [Wang+ 2019]
  - machine-in-the-loopによるデータセット改良
    - JCommonsenseQA 2.0: 計算機と人の協働による常識推論データセットの改良 [栗原+ 2023]
- より専門性の高いベンチマークの構築
  - cf. MMLU [Hendrycks+ 2021]
    - 数学、物理、法学、歴史など57ジャンルの4択問題
      - 大学院進学適性試験(GRE)、米国医師免許試験などを含む
    - GPT-3: 44% (2020) → Flan-PaLM+CoT: 75% (2022) → GPT-4: 86% (2023)
- LLMの生成性能を評価するベンチマークの検討

大規模な日本語モデル構築・共有のためのプラットフォームの形成 (相澤彰子先生代表)

- 学際大規模情報基盤共同利用・共同研究拠点(JHPCN)共同研究採択課題(2022, 2023)
  - 「データ活用社会創成プラットフォームmdx」利用
- 構築したモデル
  - nlp-waseda/gpt2-xl-japanese
    - GPT-2 1.5B: A100 x 8, 2.5か月
  - ku-nlp/deberta-v2-{base, large}-japanese [植田 JLR2023]
    - DeBERTa base: A100 x 8, 3週間
    - DeBERTa large: A100 x 8, 5週間

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

LLM 勉強会

本勉強会では、自然言語処理および計算機システムの研究者が集まり大規模言語モデルの研究開発について定期的に情報共有を行っています。

具体的には、以下の目的で活動しています。

- オープンソースかつ日本語に強い大規模モデルの構築とそれに関連する研究開発の推進
- 上記に関心のある自然言語処理および関連分野の研究者によるモデル構築の知見や最近の研究の発展についての定期的な情報交換
- データ・計算資源等の共有を前提とした組織横断的な研究者間の連携の促進
- モデル・ツール・技術資料等の成果物の公開

News

謝辞

貴重な機会を与えていただいたABCIおよび産業技術総合研究所の方々に深く感謝いたします。