2023.08.21

対話型文章生成AIの構築技術

ー 各構成技術の役割や導入背景などの解説 ー

東北大学
データ駆動科学・AI教育研究センター
大学院情報科学研究科 人工知能基礎学講座
鈴木潤

自己紹介

- 名前： 鈴木 潤

専門分野：人工知能，機械学習，自然言語処理

- 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
- 2020.07 - 現在　東北大学 データ駆動科学・AI教育研究センター 教授
  大学院情報科学研究科 人工知能基礎学講座 (協力講座)
  (研究室は2021.4から)
- 2020.04 - 2022.04　Google LLC (Visiting Researcher)

主に計算機が言語を効率的に学習する方法についての研究に従事

講演

NIIオープンハウス 2023 基調講演

https://www.nii.ac.jp/event/openhouse/2023/

一般向け（非専門家向け）

言語処理学会第29回年次大会 (NLP2023)
緊急パネル：ChatGPTで自然言語処理は終わるのか？
https://www.anlp.jp/nlp2023/#special_panel

講演

PAKDD 2023 Tutorial
https://pakdd2023.org/tutorials/

非NLP研究者向け

概要

- ChatGPTに代表される対話型文章生成AIが，研究分野のみならず産業界や一般社会においても注目を集めている．対話型文章生成AIはニューラルネットによる言語モデルを基本（出発点）として幾つかの要素技術の集合体として構築されている．本講演では，対話型文章生成AIを構築する際に用いられる要素技術の集合をおおまかに整理して紹介する．また，各技術の役割や導入背景などを説明し，どのような過程を経て現在の対話型文章生成AIに到達したのかについて説明する

[注意事項] 事前のお断り

- 本資料には講演者独自の解釈が含まれています
  - ここに書かれていることは必ずしも一般的に普及している知見ではない場合があります
  - 議論や洞察を深めるためのきっかけとしてあえて書いている部分もあります

- 本資料の情報の正確性について
  - 対話型文章生成AI / ChatGPT 関連の情報は頻繁に更新
  - あくまで2023.08.21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

対話型文章生成AI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

現在の代名詞：ChatGPT

[参考] 年単位のNLP分野の重要技術

2013 Neural word embeddings / 単語分散表現
2014 Neural encoder-decoder / 系列変換器
2015 Attention mechanism / 注意機構
2016 Subword / サブワード分割
2017 Transformer / 自己注意型ネットワーク
2018 Neural language models / ニューラル言語モデル
2019 Masked neural LMs / マスク型言語モデル
2020 Large LMs: GPT-3 / 巨大言語モデル
2021 Prompt tuning, engineering / プロンプト開発
2022 Instruction (Chat) tuning / 言語生成品質の向上
2023 (???)

今日の話題

- 文章生成AI (言語モデル) がうまくいく第一要因は？
- 言語モデルに Transformer は たまたま？
- プロンプト (汎用/非特化型) のきっかけは？
- (LLM構築の取り組み)

- (昨今の)「言語モデル」が文章生成AIとしてうまく機能している第一要因は？

[参考] ChatGPTを構成する技術

- 基盤
  - 言語モデル

- 成功の要因
  - ① 多層ニューラルネットワーク (DNN) の利用
  - ② 大規模化 (パラメタ数/データ量)
  - ③ 指示文設計 (プロンプトエンジニアリング)
  - ④ 指示文 (+対話文) チューニング
    - ⑤ 人手点数付け結果の活用

[参考] ChatGPTを構成する技術

TOHOKU UNIVERSITY

2階部分 (指示文の活用/学習)

指示文設計
指示文 (+対話文) 微調整学習
人手点数付けの利用

獲得するスキル?

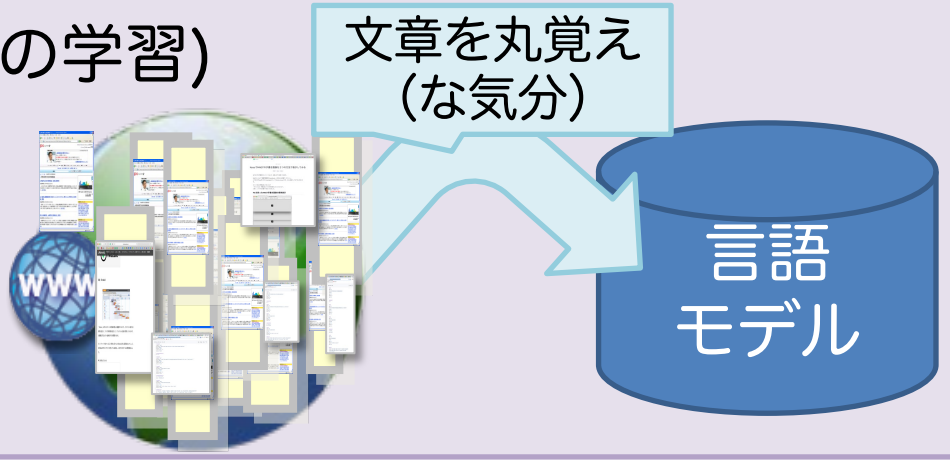
指示文の理解

対話的なやりとりの理解

不適切発言の抑制

1階部分 (言語モデルの学習)

ニューラルネット
巨大言語モデル
大規模データ
(教師あり学習，確率モデル)



文章の流暢さ

世界の知識

基盤：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
・・・

$P(Y = \$\text{BOS}\$ \text{今日 は 良い 天気 です 。} \$\text{EOS}\$)$ 高

$P(Y = \$\text{BOS}\$ \text{良い です 。天気 は 今日} \$\text{EOS}\$)$ 低

言語
モデル

確率は大量のデータから何かしらの形で計算

言語モデルの例: n-gram 言語モデル

- Example: Google n-gram

BLOG ›

All Our N-gram are Belong to You

THURSDAY, AUGUST 03, 2006

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of *one trillion words* from public Web pages.

We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

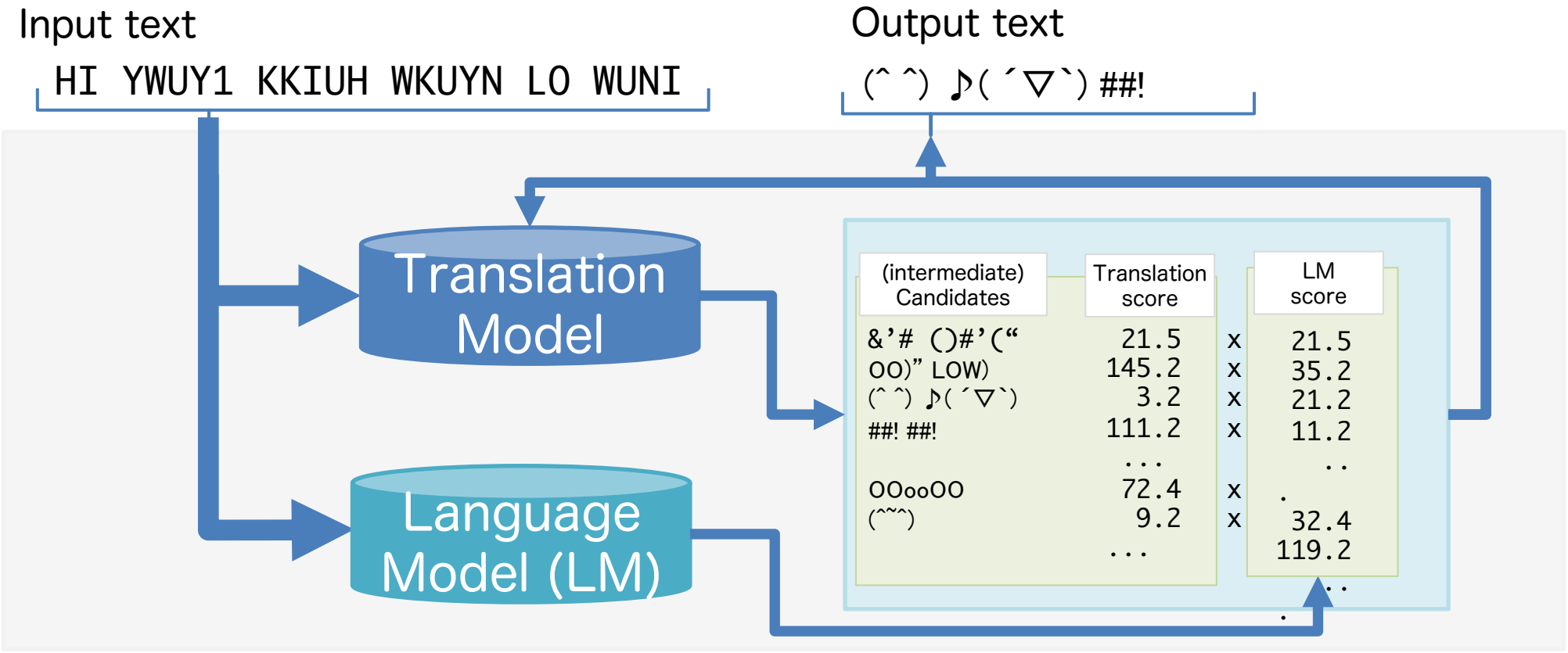
https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

Context $Y_{<j}$	Target word y_j	Count	Probability
serve as the	incoming	92	0.00049
serve as the	independent	794	0.00423
serve as the	index	223	0.00119
serve as the	indication	72	0.00038
serve as the	indicator	120	0.00064
serve as the	indicators	45	0.00024
serve as the	indispensable	111	0.00059
serve as the	indispensible	40	0.00021
serve as the	individual	234	0.00125
serve as the	industrial	52	0.00028
serve as the	industry	607	0.00324
serve as the	info	42	0.00022
serve as the	informal	102	0.00054
serve as the	information	838	0.00447
serve as the	informational	41	0.00022
serve as the	initial	5331	0.02843
serve as the	initiating	125	0.00067
serve as the	initiation	63	0.00034
serve as the	initiator	81	0.00043
serve as the	.....		
		Total	187491

昔の言語モデルの使い方

- モデルが生成した候補文の尤もらしさを計測

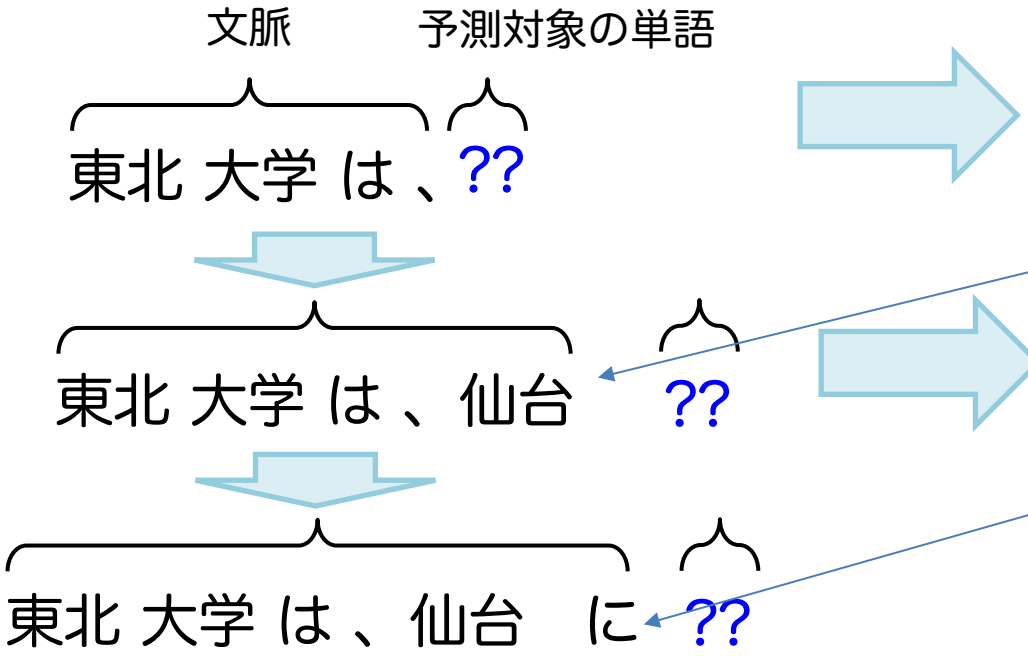
例：統計翻訳 (Statistical Machine Translation: SMT)

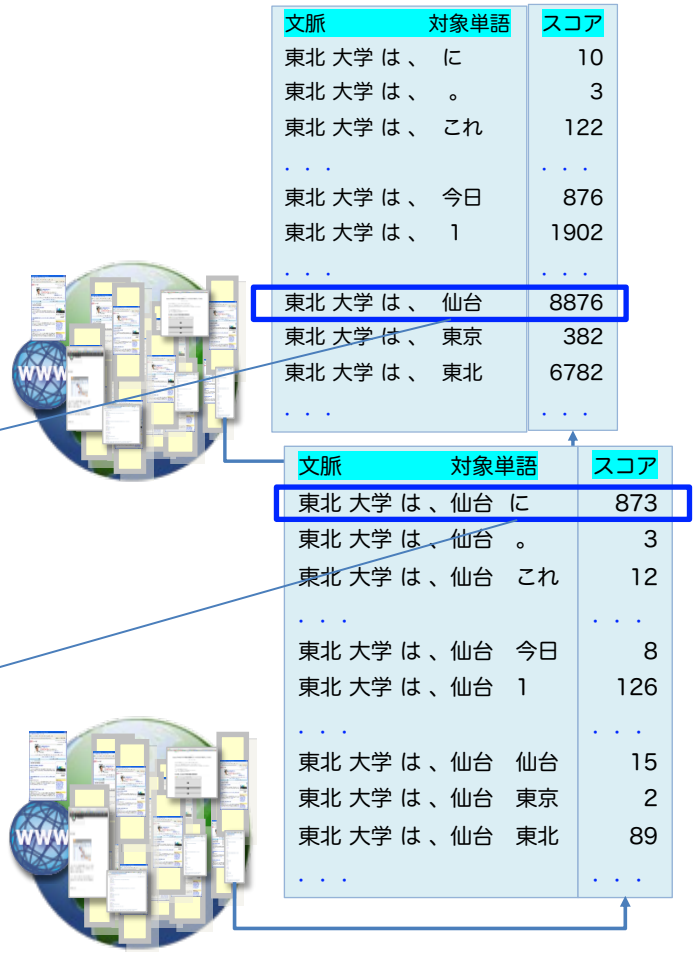


言語モデルによる文生成

- 文章の先頭から1単語ずつ次の単語を予測

  => 予測した単語は文脈として再利用
  生成が終わるまで繰り返す





文脈	対象単語	スコア
東北 大学 は 、	に	10
東北 大学 は 、	。	3
東北 大学 は 、	これ	122
東北 大学 は 、	今日	876
東北 大学 は 、	1	1902
東北 大学 は 、	仙台	8876
東北 大学 は 、	東京	382
東北 大学 は 、	東北	6782

文脈	対象単語	スコア
東北 大学 は 、仙台	に	873
東北 大学 は 、仙台	。	3
東北 大学 は 、仙台	これ	12
東北 大学 は 、仙台	今日	8
東北 大学 は 、仙台	1	126
東北 大学 は 、仙台	仙台	15
東北 大学 は 、仙台	東京	2
東北 大学 は 、仙台	東北	89

言語モデルの機能というよりは利用例の一つ?

ニューラル言語モデル

DNN: deep neural network

- 言語モデル (確率モデル) をDNNで近似
  - [文脈] => [正解次単語] を分類問題として学習
  - 文章中の全単語を対象として学習

Web上にある大量のテキストなどを利用

東北大学は、建学以来の伝統である「研究第一」と「門戸開放」の理念を掲げ、．．．

[文脈] => [正解次単語]

[東北 大学 は 、] => [建学]
[東北 大学 は 、 建学] => [以来]
[東北 大学 は 、 建学 以来] => [の]
[東北 大学 は 、 建学 以来 の] => [伝統]
．．．

東北大学は、1907年（明治40年）に東北帝国大学として創立し、当初から、．．．

[文脈] => [正解次単語]

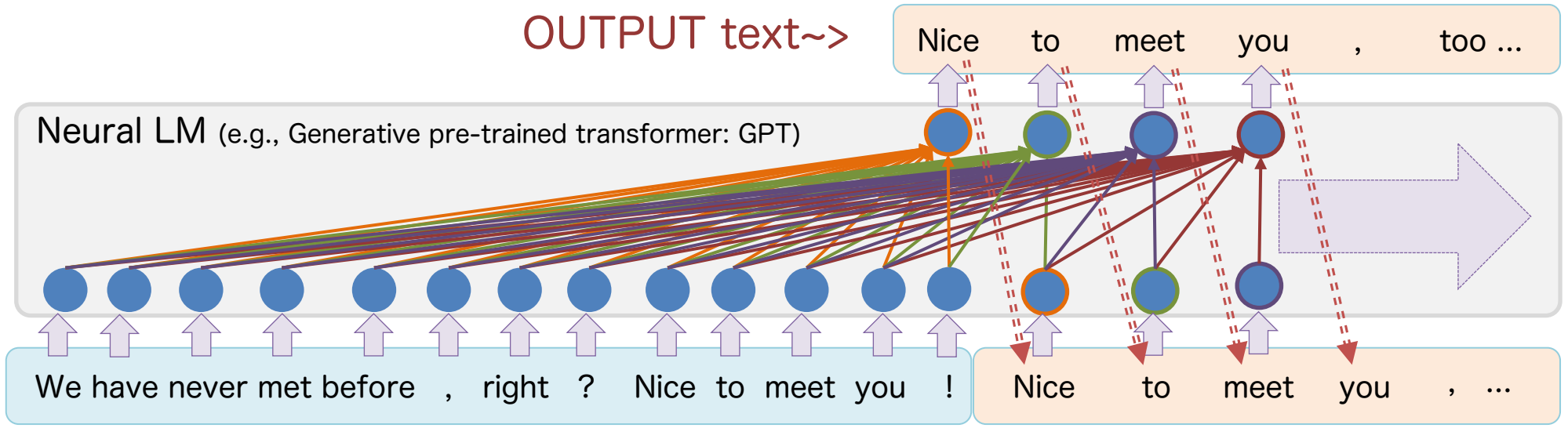
[東北 大学 は 、] => [1]
[東北 大学 は 、 1] => [9]
[東北 大学 は 、 1 9] => [0]
[東北 大学 は 、 1 9 0] => [7]
．．．

ニューラル言語モデルの学習データ

ニューラル言語モデル

$$\underset{\boldsymbol{y}_j}{\text{argmax}}\ P_{\boldsymbol{\theta}}(\boldsymbol{y}_j \mid \boldsymbol{Y}_{<j})$$

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
,	,	,	,
.	.	.	.



文章生成AIがうまくいく第一要因は？

- 成功の要因は「言語モデル」の能力/機能？
  - 言語モデルはただの確率モデルのはず
  - n-gram LMをどんなに巨大化してもChatGPTにはならない

文章生成AIがうまくいく第一要因は？

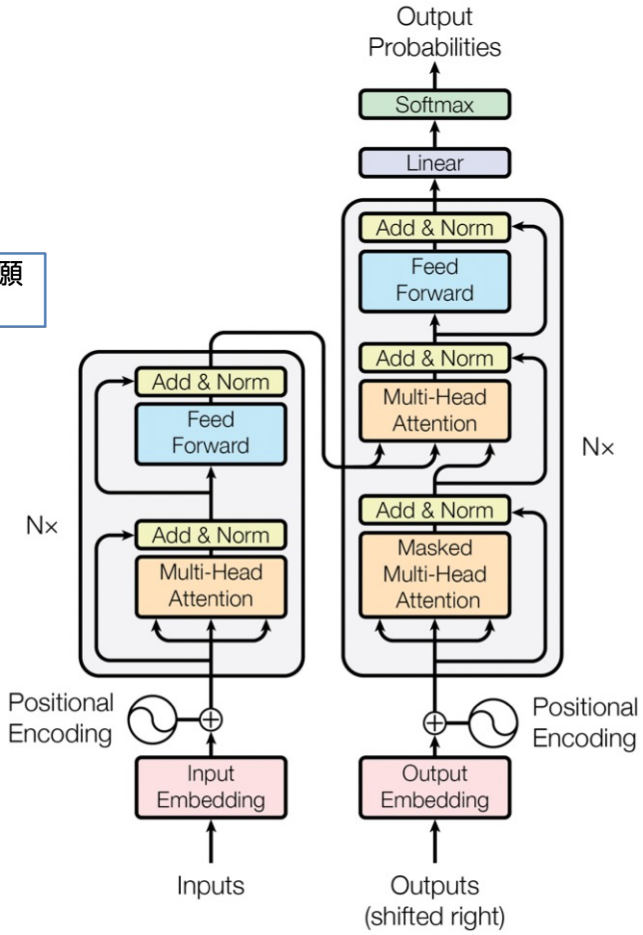
- 成功の要因は「言語モデル」の能力/機能？
  - 言語モデルはただの確率モデルのはず
  - n-gram LMをどんなに巨大化してもChatGPTにはならない

- 何が実際の要因なのか？
  - 言語（単語）の意味や使われ方の観点で似ている/似ていないが判断できるようになったこと
    => 分散表現 (単語埋め込み) でも語られていた能力
  - 言語の「世界モデル」として覚えておくのに適しているモデルだったこと
    - Transformerだったのも運が良かった？

- 言語モデルに Transformer はたまたま？

[参考] Transformer

- 深層ニューラルネットワークの種類の一つ
  - 2017年：主に機械翻訳用として提案
    翻訳関連で最初に流行った

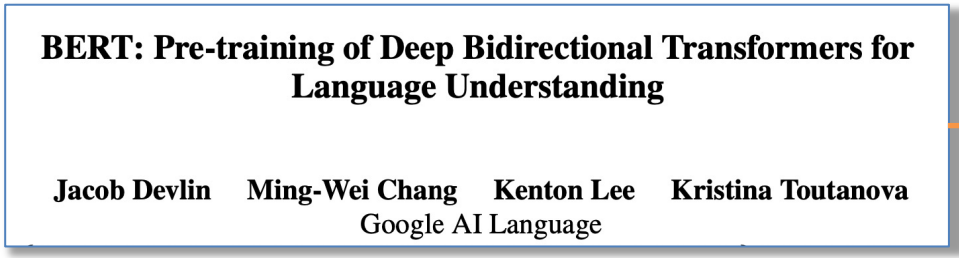


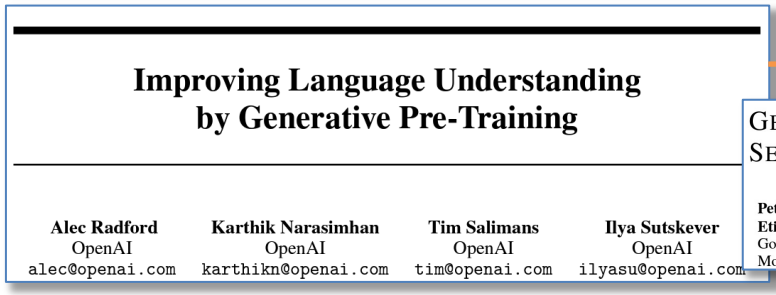
https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

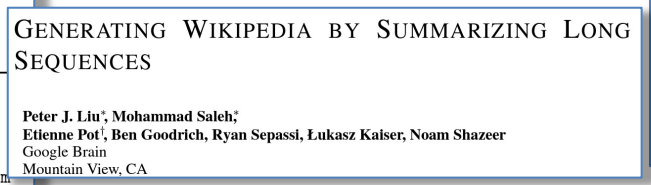
- ニューラルネットに基づくAI技術において最重要技術の一つ

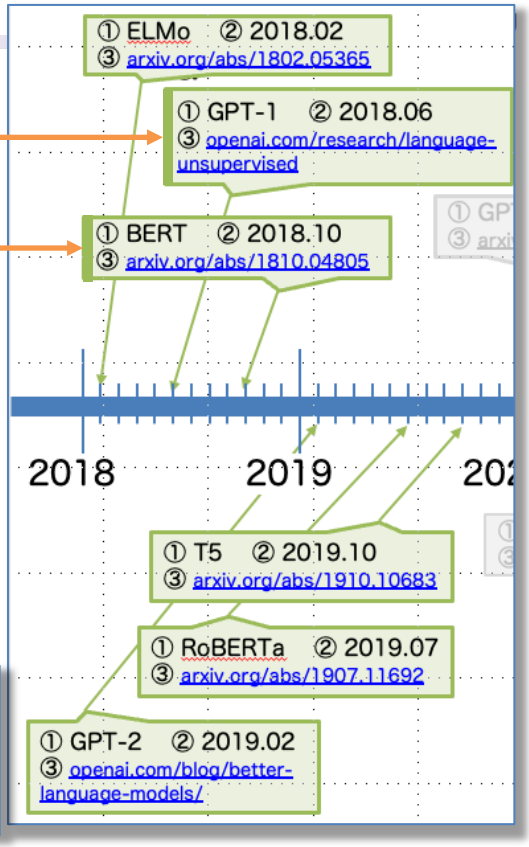
[参考] Transformer

- 2018: 言語モデルBERT (とGPT) のベースモデルとして採用
  - => 言語の基盤モデルに







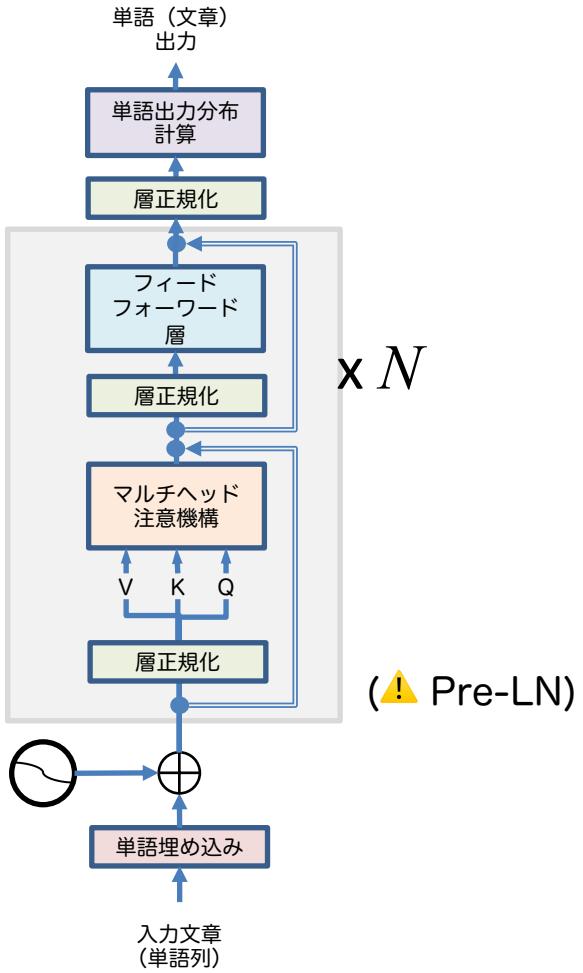


- 2023有名なニューラル言語モデルのほぼ全てで採用
  - 例：GPT-3 (ChatGPT/GPT-4), PaLM, LLaMA, OPT, BLOOM, ...
- 画像/音声/信号処理などでも広く普及

Transformer は言語モデルに適している？

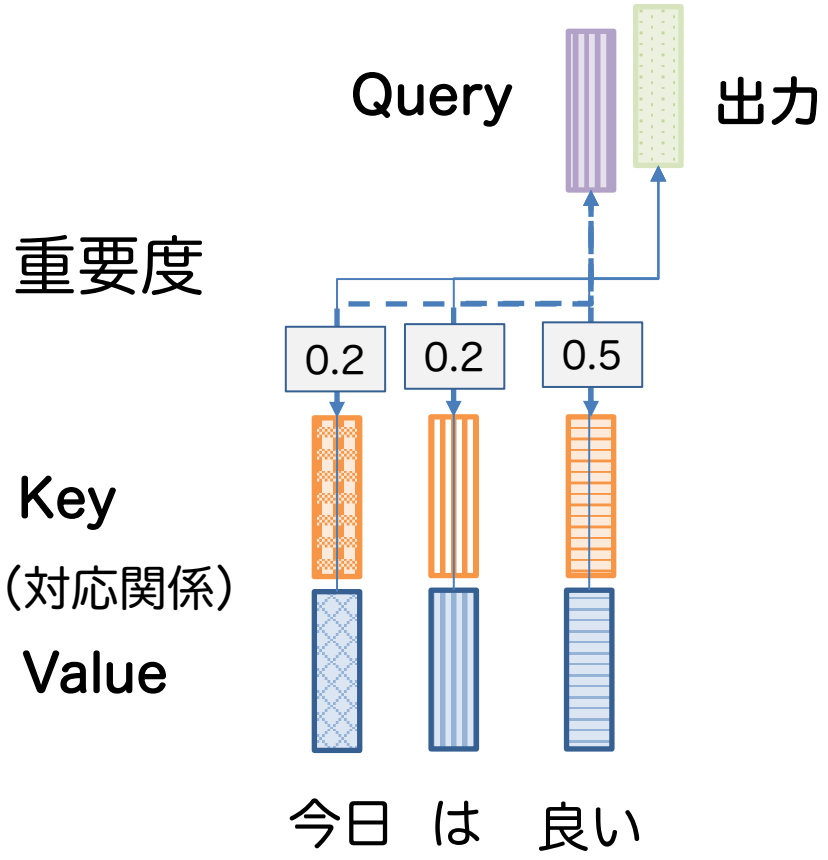
- 主にAttentionとFFNNで構成されている

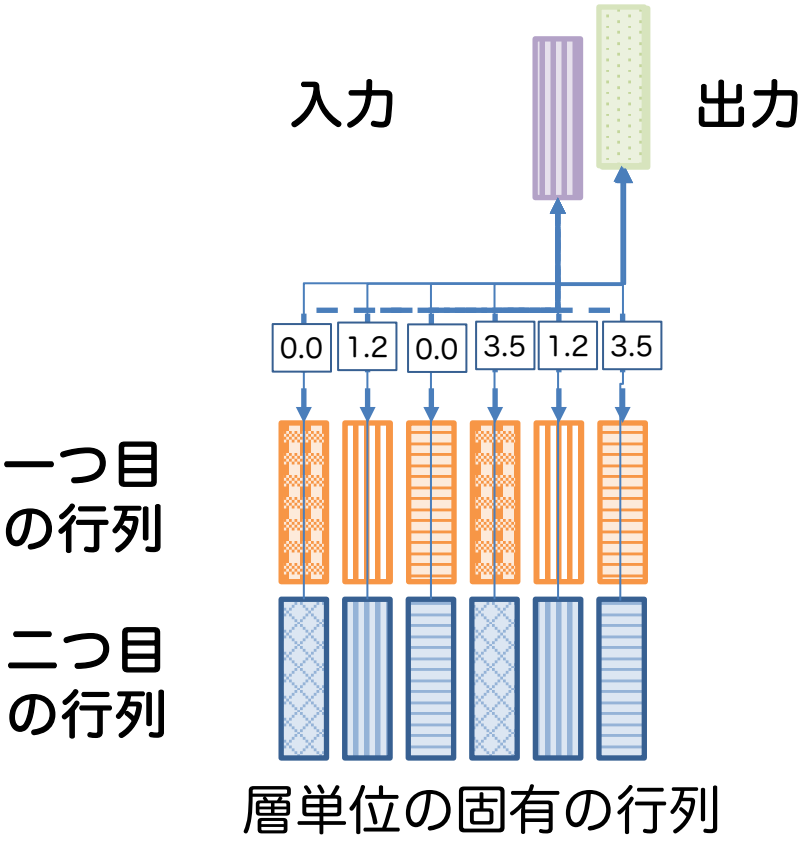
GPT-3		12,288				
Encoder	in	out	bias	sub total	num	total
embedding	50,257	12,288	0	617,558,016	1	617,558,016
Position	12,288	4,096	0	50,331,648	1	50,331,648
k	12,288	12,288	12,288	151,007,232	96	14,496,694,272
v	12,288	12,288	12,288	151,007,232	96	14,496,694,272
q	12,288	12,288	12,288	151,007,232	96	14,496,694,272
out	12,288	12,288	12,288	151,007,232	96	14,496,694,272
layernorm1	1	12,288	12,288	24,576	96	2,359,296
ff1	12,288	49,152	49,152	604,028,928	96	57,986,777,088
ff2	49,152	12,288	12,288	603,992,064	96	57,983,238,144
layernorm2	1	12,288	12,288	24,576	96	2,359,296
layernorm_last	1	12,288	12,288	24,576	1	24,576
						174,629,425,152



Transformer は言語モデルに適している？

- Attention / FFNN 共にメモリとその呼び出し処理と見做せる？





Transformer は言語モデルに適している？

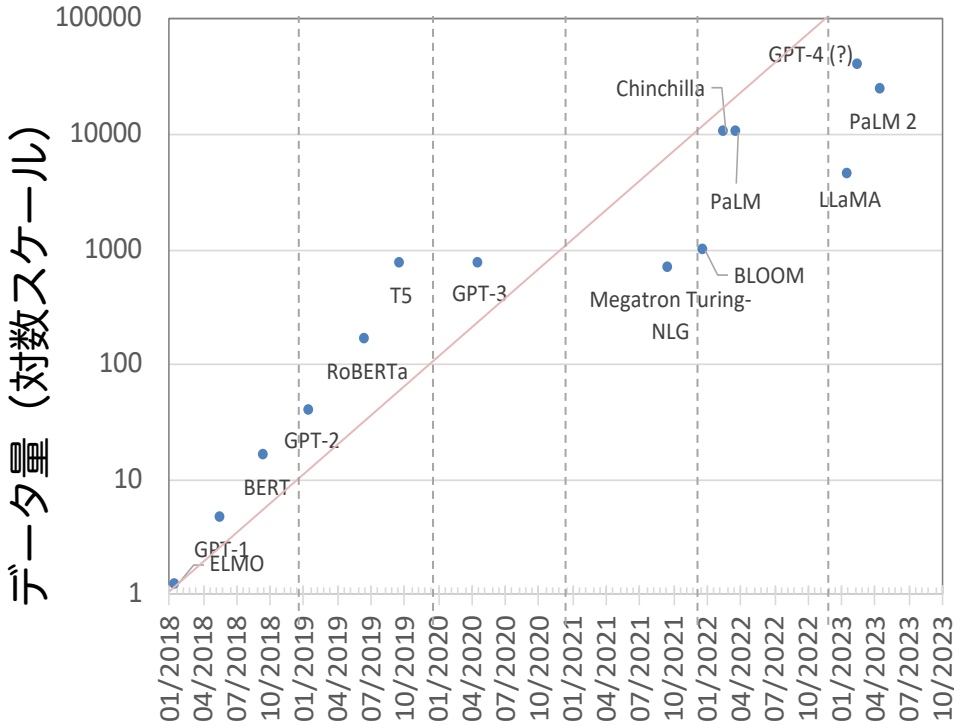
- Attention / FFNN 共にメモリとその呼び出し処理と見做せる？
  - 学習データに過適応できる性質
  - 学習はなるべく手を抜く性質
    - Shortcut learning, superficial cues
    - (この辺りの性質が構造的な情報の学習をうまく実現する要因?)
  - 「まる覚えする能力」が高いと想定できる
    - ただし，「文章」を丸覚えしているわけではなくベクトル列の情報を丸覚えしている（しようとしている）

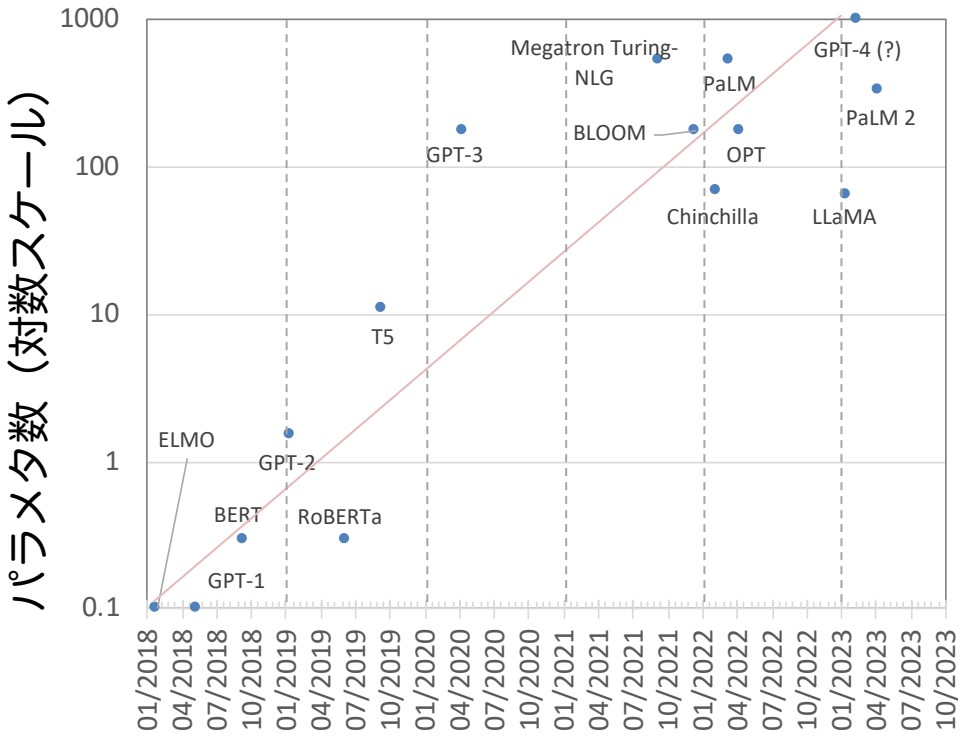
年毎の大規模化

- 学習データ量 / パラメタ数：年々順調に増加

	トークン数	ファイル容量
GPT-2:	0.01 B	0.04 TB
GPT-3:	0.5 B	0.75 TB
GPT-4:	13 T	20 TB (憶測)

	パラメタ数	必要メモリ量
GPT-2:	1.5 B	6 GB
GPT-3:	175 B	700 GB (ChatGPT)
GPT-4:	1,800 B	7,200 GB (憶測)



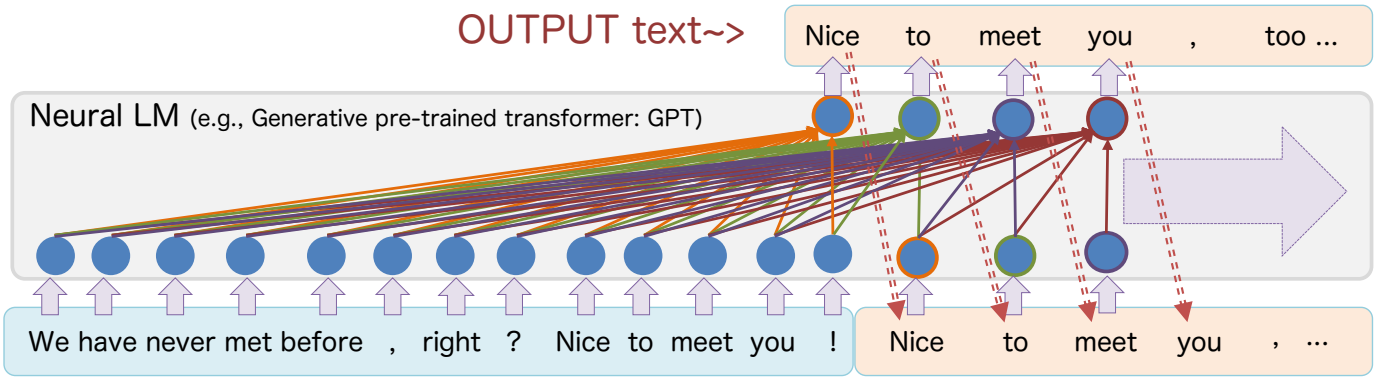


- プロンプト (汎用/非特化型) のきっかけは？

指示文 (プロンプト) とは？

文脈 = 指示文 (プロンプト) と考えてよい

- 言語モデル：与えられた文脈に基づいて文章の続きを生成 (文章の補完)

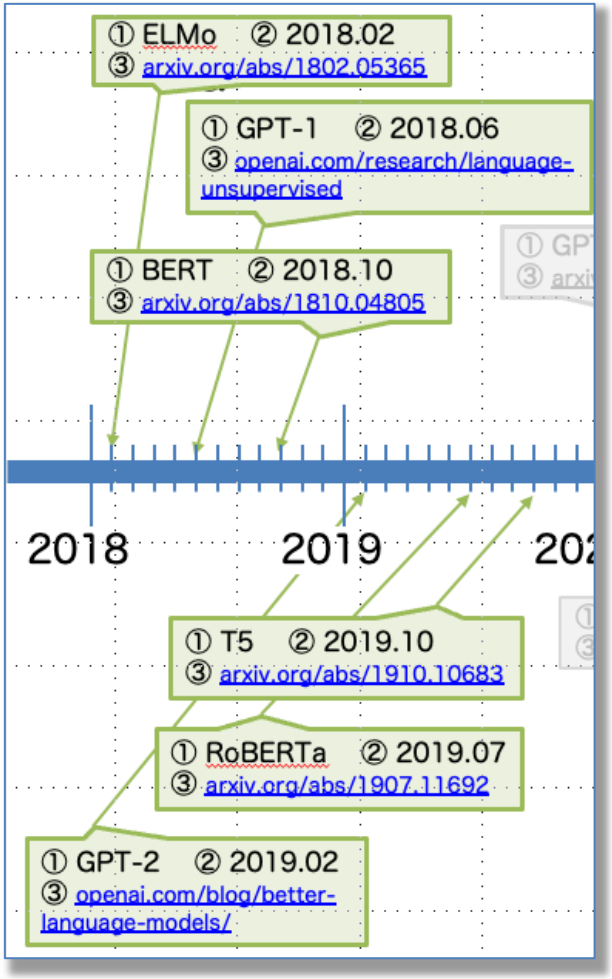


- 文脈を工夫すると様々な文章を生成することが判明
- なぜ特別な用語を用いるのか

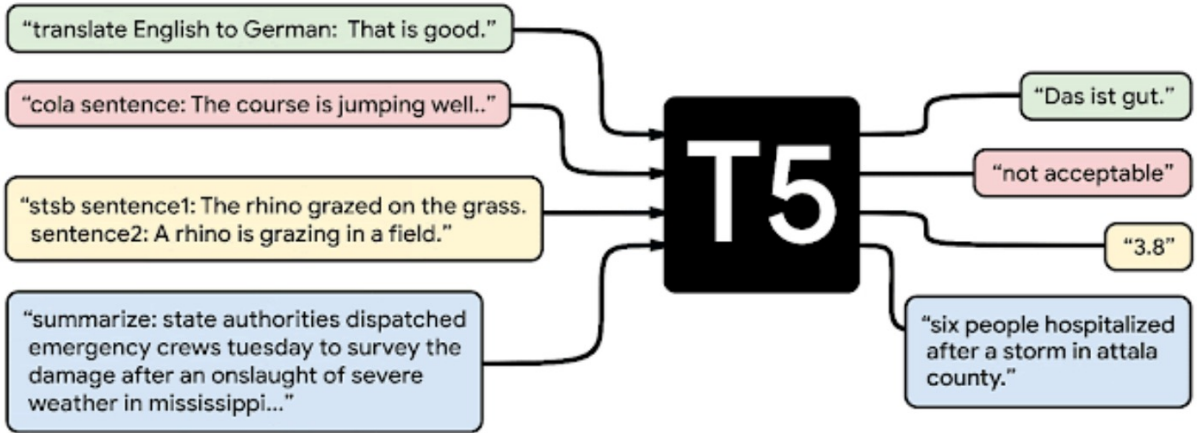
  => 単に「文脈」と呼ぶ以上の何か

[参考] 言語処理タスクの解き方

- 従来：それぞれのタスクに特化したモデル/方法の構築
  - 特定のタスクのみ効果的に解ける
- 一つのモデルであらゆる言語処理タスクを解くことを目指す => 生成モデル
  - 現在の対話型文章生成AIの先駆け (T5)



例：



Copied from https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

指示文設計 (プロンプトエンジニアリング)

- GPT-3: 指示文 (プロンプト) の考え方の転換点
  - Finetuningをしない / その代わりに事例を使う

Zero-shot

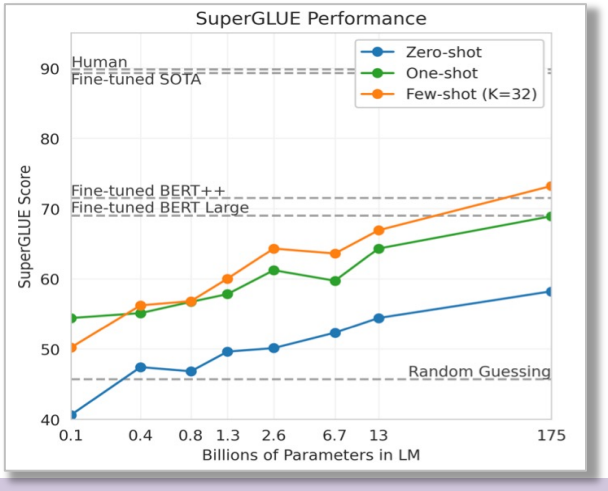
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```



Language Models are Few-Shot Learners

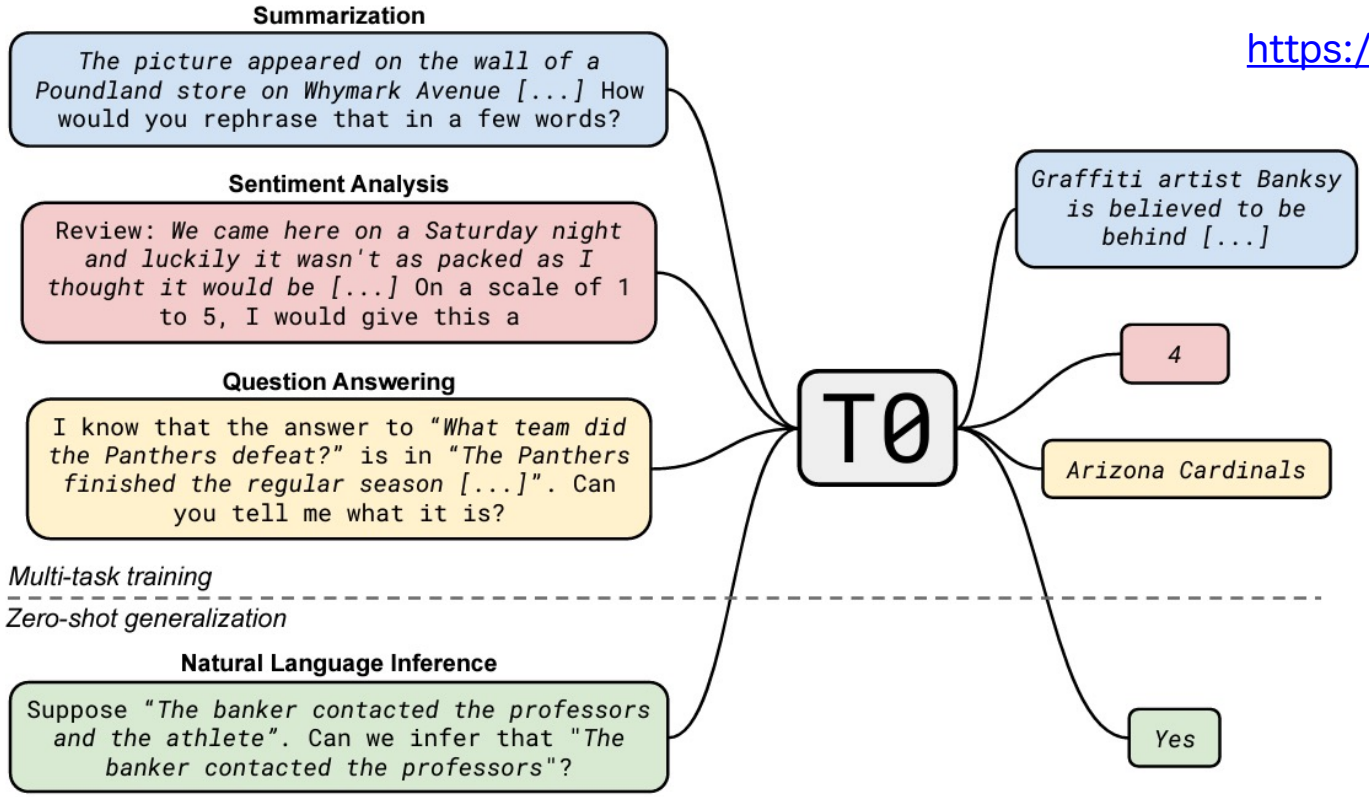
https://arxiv.org/abs/2005.14165

指示文設計 (プロンプトエンジニアリング)

- T0: 指示文をそのまま利用

Multitask Prompted Training Enables Zero-Shot Task Generalization

https://arxiv.org/abs/2110.08207

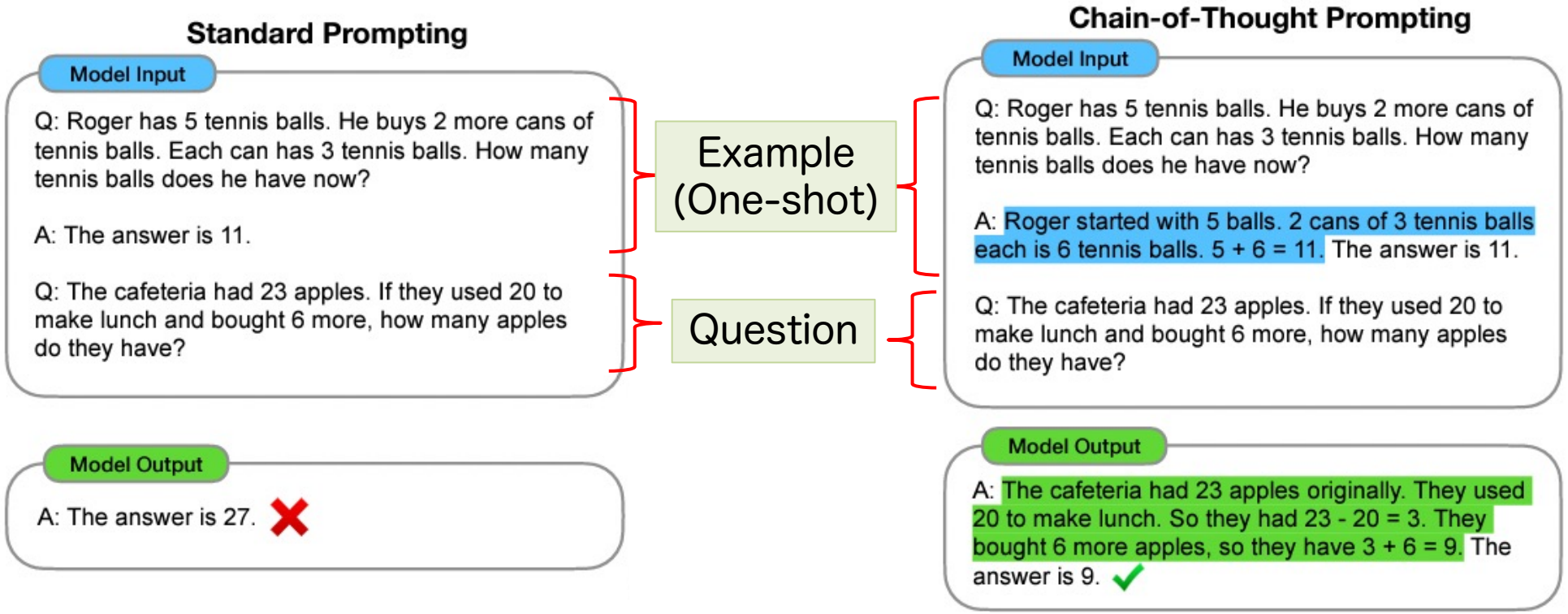


おおむね現在使われるプロンプトに近い状態

指示文設計 (プロンプトエンジニアリング)

- Chain-of-thought: 解き方を教える

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
https://arxiv.org/abs/2201.11903



まとめ

- プロンプト発展の経緯
  => 一つのモデルで多数のタスクに対応したい

- 時代とともに微妙に「プロンプト」が指すものが広がっている（気がする）

- (LLM構築の取り組み)

llm-jpの活動

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

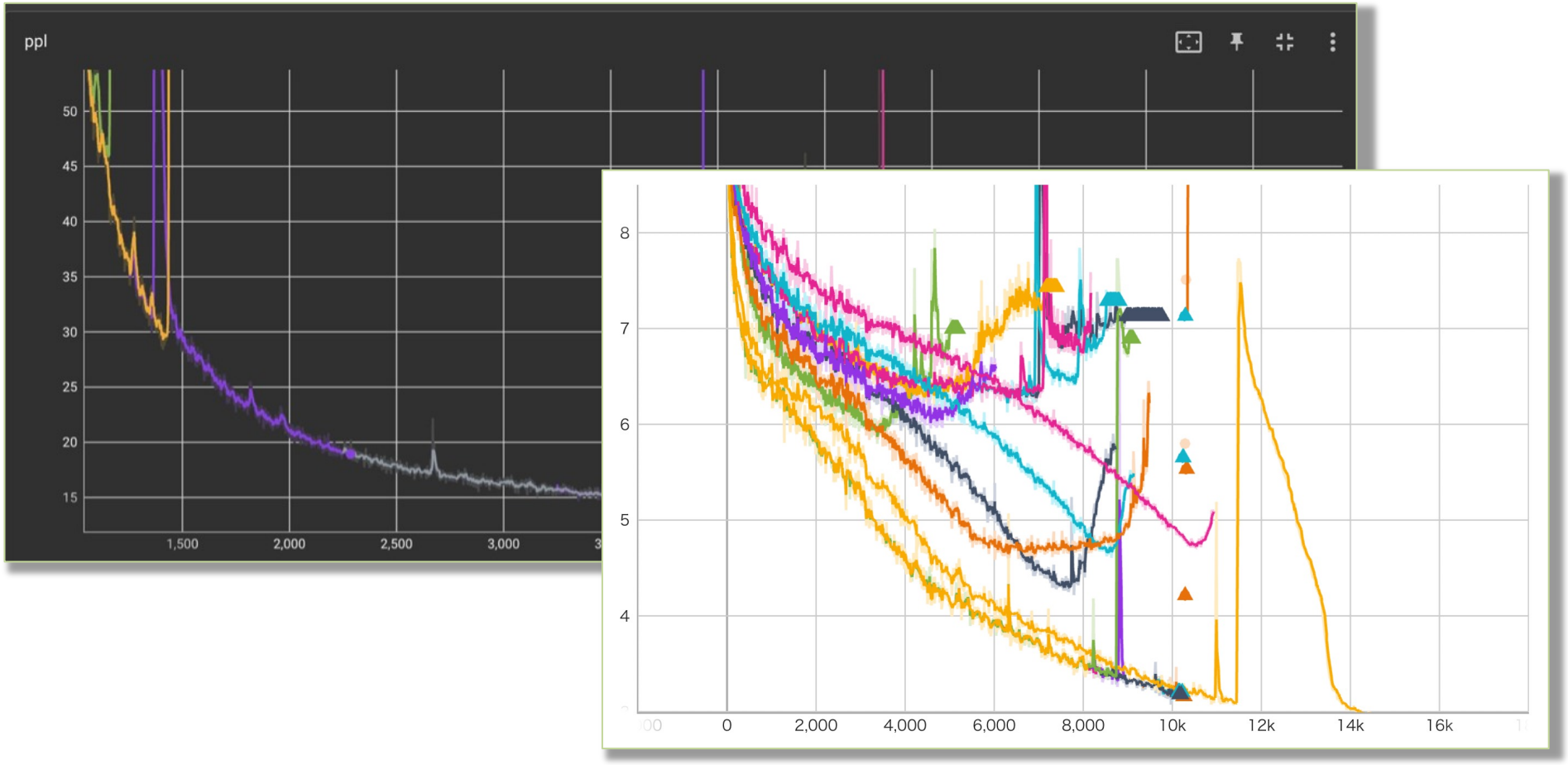
- LLM 勉強会
  - 本勉強会では、自然言語処理および計算機システムの研究者が集まり大規模言語モデルの研究開発について定期的に情報共有を行っています。
- 具体的には、以下の目的で活動しています。
  - オープンソースかつ日本語に強い大規模モデルの構築とそれに関連する研究開発の推進
  - 上記に関心のある自然言語処理および関連分野の研究者によるモデル構築の知見や最近の研究の発展についての定期的な情報交換
  - データ・計算資源等の共有を前提とした組織横断的な研究者間の連携の促進
  - モデル・ツール・技術資料等の成果物の公開

モデル構築検討WG幹事として活動

LLMの構築

- LLMの学習はそれほど単純ではない

https://github.com/facebookresearch/metaseq/blob/main/projects/OPT/chronicles/10_percent_update.md



https://huggingface.co/blog/bloom-megatron-deepspeed#bf16optimizer

- まとめ

まとめ (1/2)

- ChatGPT (対話型文章生成AI) とは？

  => 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

- 技術
  - 基盤：言語モデル
  - ① 深層ニューラルネットワーク (DNN) の利用
  - ② 大規模化 (パラメタ数/データ量)
  - ③ 指示文設計 (プロンプトエンジニアリング)
  - ④ 指示文 (+対話文) チューニング
    - ⑤ 人手点数付け結果の活用

まとめ (2/2)

- 本日の話題
  - 文章生成AI (言語モデル) がうまくいく第一要因は？
  - 言語モデルに Transformer はたまたま？
  - プロンプト (汎用/非特化型) のきっかけは？
  - (LLM構築の取り組み)