※図表は論文・著者スライドより引用

Knowledge-in-Context: Towards Knowledgeable Semi-Parametric Language Models

著者：Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, Jianshu Chen*
（Tencent AI Lab, Bellevue, WA 98004, USA）
@ICLR2023
読み手：井田龍希（豊田工業大学 知能数理研究室 M2）

本論文を選んだ理由

- 事前学習中, fine-tuning中に外部知識を言語モデルに組み込む研究が多数
  - 場合によっては, 外部知識が性能向上に寄与しない

    適切な外部知識を利用できていない？そもそも外部知識が有用でない？

- **本論文：複数の外部知識源から関連する外部知識源を動的に選択・利用**
  - 知識源の選択はインスタンスごと, 外部知識を使わないという選択も

    ⇒ **適切な外部知識の利用方法**に踏み込んでいる

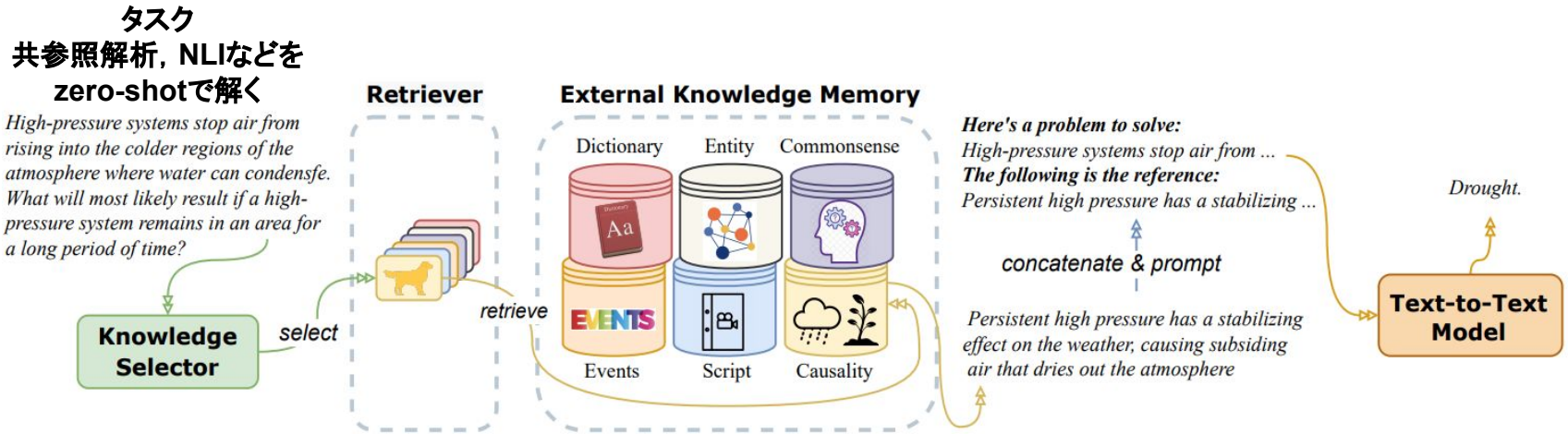
今後, 言語モデルにおける外部知識のより効果的な利用を目指す上で

重要な研究なのでは？

まとめ

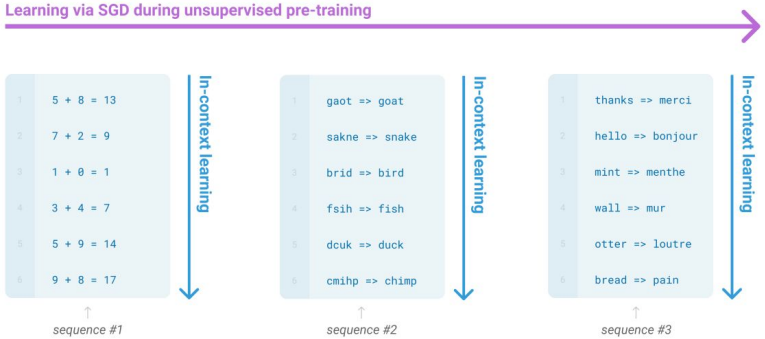
※ 言語モデルのアーキテクチャの変更などの話ではないです.

- **複数の外部知識源**を利用したRetrieval-based言語モデル
  - インスタンスごとに適切な外部知識源を動的に選択
  - 複数のタスクで外部知識 + 入力文から回答をinstruction tuningで学習



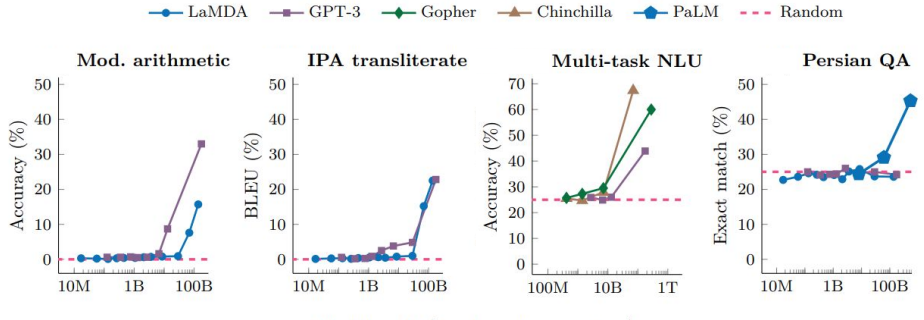
背景:In-context learning

- タスクの指示や例を与えて，複数のNLPタスクにzero/few-shotで適応
  - 一定スケール以上の膨大なパラメータを持つLLMのみ
  - 理由1:事前学習における**暗黙的**なマルチタスク学習の結果？
  - 理由2:膨大なパラメータに暗に知識を保持？(全ての知識の保持は不可能)



[Brown et al., Language Models are Few-Shot Learners, NeurIPS2020]

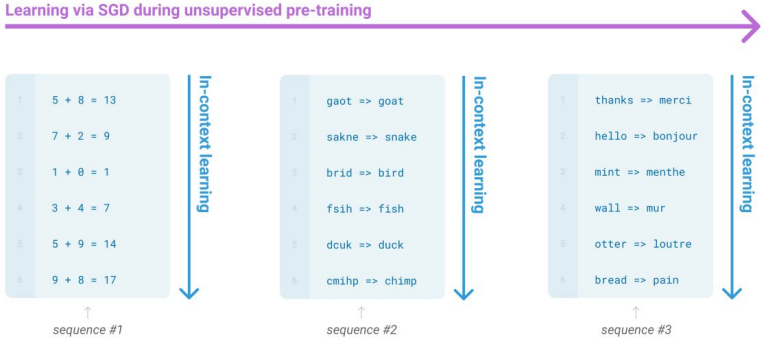
一定サイズより小さいモデル:ほぼランダムな性能
一定サイズを超えたモデル:ランダムを大幅に上回る性能



[Wei et al., Emergent Abilities of Large Language Models, TMLR2022]

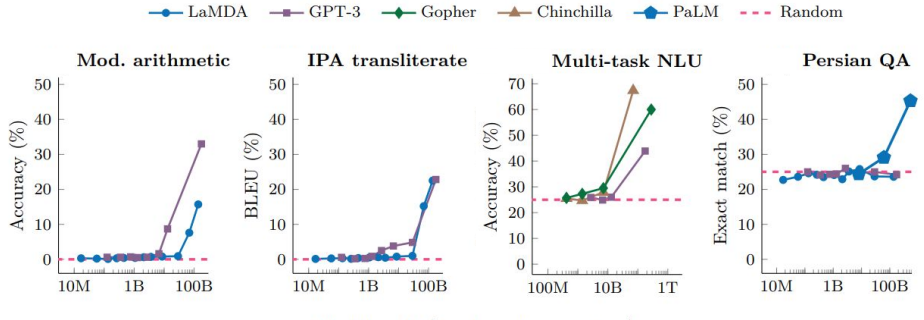
背景:In-context learning

- タスクの指示や例を与えて，複数のNLPタスクにzero/few-shotで適応
  - 一定スケール以上の膨大なパラメータを持つLLMのみ
  - 理由1:事前学習における**暗黙的**なマルチタスク学習の結果？
  - 理由2:膨大なパラメータに暗に知識を保持？(全ての知識の保持は不可能)



[Brown et al., Language Models are Few-Shot Learners, NeurIPS2020]

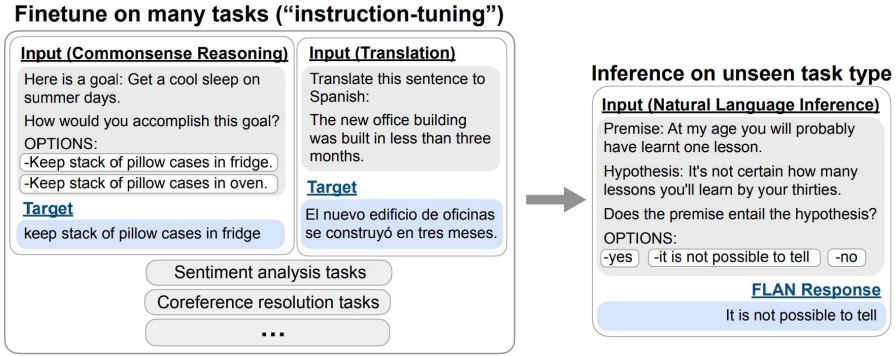
一定サイズより小さいモデル:ほぼランダムな性能
一定サイズを超えたモデル:ランダムを大幅に上回る性能

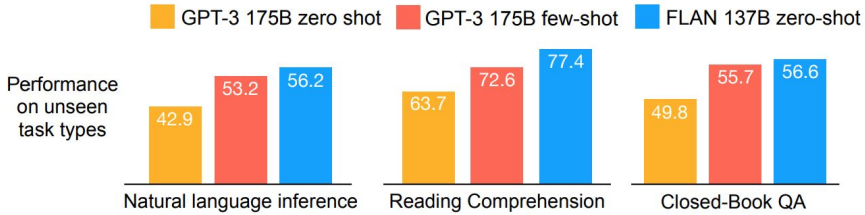


[Wei et al., Emergent Abilities of Large Language Models, TMLR2022]

背景:Instruction tuning

- **明示的に**タスクの指示や例を含むプロンプトによるfine-tuning
  - タスクの指示や例から出力へのマッピングを学習
  - より少ないパラメータのモデルで高いzero-shot性能を達成

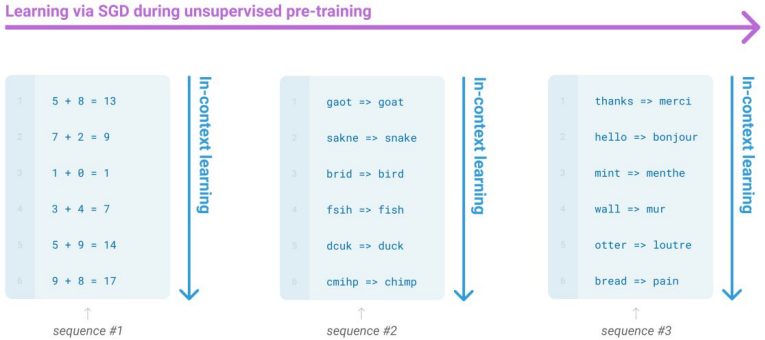




[Wei et al., Finetuned Language Models Are Zero-Shot Learners, ICLR2022]

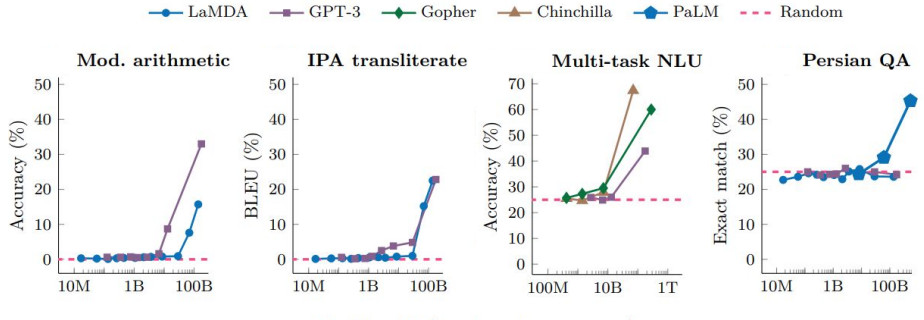
背景:In-context learning

- タスクの指示や例を与えて，複数のNLPタスクにzero/few-shotで適応
  - 一定スケール以上の膨大なパラメータを持つLLMのみ
  - 理由1:事前学習における**暗黙的**なマルチタスク学習の結果？
  - 理由2:膨大なパラメータに暗に知識を保持？(全ての知識の保持は不可能)



[Brown et al., Language Models are Few-Shot Learners, NeurIPS2020]

一定サイズより小さいモデル:ほぼランダムな性能
一定サイズを超えたモデル:ランダムを大幅に上回る性能



[Wei et al., Emergent Abilities of Large Language Models, TMLR2022]

背景:Retrieval-based Language Model

- 外部のテキストコーパスなどから関連知識を検索して利用
  - **利点1:知識に基づく回答**が可能, Hallucinationを軽減
  - 利点2:外部のテキストコーパスを変更するだけで**新しい知識への更新**が可能

[Ram et al., In-Context Retrieval-Augmented Language Models, TACL, 2023]

既存手法と本論文の手法の違い

- 既存手法：外部知識源として単一のテキストコーパスを利用

  有用な知識を含む文が少なく，検索が困難

- 本論文：外部知識源として6種類の構造化，半構造化された知識源を利用

  知識が密，簡潔に記述．複数の知識源により適切な知識の利用が可能に

  - instruction tuningによって統一的なモデルで複数タスクでの外部知識利用

Wikipedia（テキストコーパス）vs CausalBank（構造化された知識源）

Q. 高気圧が長期間にわたってある地域に留まる場合，最もありそうな結果は？ A. 干ばつ
Wikipedia（テキストコーパス）　**関連はあるが，有用な知識とは言えない** 高気圧はanticyclonesとも呼ばれる．英語の天気図では，高気圧中心は英語のHの文字で識別される．
CausalBank（因果関係に関する構造化された知識源） 持続的な高気圧は天気に安定的な効果，result in，下降気流を引き起こし，大気を乾燥させる． **< Subject　Relation　Object >**

既存手法と本論文の手法の違い

- 既存手法：外部知識源として単一のテキストコーパスを利用
  有用な知識を含む文が少なく，検索が困難

複数の知識源は必要？

- 本論文：外部知識源として6種類の構造化，半構造化された知識源を利用
  知識が密，簡潔に記述．複数の知識源により適切な知識の利用が可能に
  - instruction tuningによって統一的なモデルで複数タスクでの外部知識利用

Wikipedia（テキストコーパス）vs CausalBank（構造化された知識源）

Q. 高気圧が長期間にわたってある地域に留まる場合，最もありそうな結果は？
A. 干ばつ

Wikipedia（テキストコーパス）　関連はあるが，有用な知識とは言えない
高気圧はanticyclonesとも呼ばれる．英語の天気図では，高気圧中心は英語のHの文字で識別される．

CausalBank（因果関係に関する構造化された知識源）
持続的な高気圧は天気に安定的な効果，result in，下降気流を引き起こし，大気を乾燥させる．
< Subject　Relation　Object >

事前実験：異なる知識源の効果の違いを確認

知識源の種類

Category	Task	Task Reference	None	ENT	DIC	COM	EVT	SCR	CAU
Coreference	WSC	Levesque et al. (2012)	60.3	49.5	62.0	53.1	**64.7**	62.0	63.4
	Wino. debiased	Sakaguchi et al. (2021)	**59.1**	53.5	57.3	56.9	58.3	58.5	54.1
	Wino. xl	Sakaguchi et al. (2021)	63.5	63.0	64.2	**64.5**	64.3	63.8	63.5
NLI	CB	De Marneffe et al. (2019)	87.5	87.5	85.9	87.5	85.9	**90.6**	84.4
	RTE	Wang et al. (2019)	77.1	76.2	79.0	**79.2**	76.6	76.9	71.3
Paraphrase	MRPC	Dolan & Brockett (2005)	82.9	80.5	**87.7**	77.9	84.9	84.4	82.0
	QQP	Wang et al. (2018)	89.4	89.1	**89.5**	89.5	89.2	89.3	89.4
	PAWS	Zhang et al. (2019a)	**94.6**	94.2	94.3	94.4	94.4	94.5	94.2
Closed QA	ARC-Easy	Clark et al. (2018)	52.8	52.6	53.1	51.7	56.1	51.7	**64.6**
	ARC-Challenge	Clark et al. (2018)	30.9	36.2	30.9	33.5	34.2	37.2	**39.5**
	WikiQA	Yang et al. (2015)	**96.2**	95.6	95.8	95.9	95.7	95.7	96.2
Extr. QA	ReCoRD	Zhang et al. (2018)	53.9	53.9	53.2	54.0	**54.1**	53.9	53.5
Multi QA	CoS-E v1.11	Rajani et al. (2019)	60.6	**61.2**	59.9	60.8	60.1	59.7	61.1
	CosmosQA	Huang et al. (2019)	69.1	69.0	68.3	**69.7**	67.9	67.7	66.4
	DREAM	Sun et al. (2019)	62.4	63.8	63.5	62.5	63.3	**63.8**	62.7
	OpenBookQA	Mihaylov et al. (2018)	56.2	54.7	54.7	57.4	**58.2**	55.7	57.6
	PIQA	Bisk et al. (2020)	71.7	72.5	71.6	71.5	71.5	70.6	**74.3**
	QASC	Khot et al. (2020)	97.8	**98.1**	98.0	98.0	98.1	97.8	97.6
	QuAIL	Rogers et al. (2020)	68.3	68.3	72.9	**73.5**	72.9	66.6	68.6
	QuaRTz	Tafjord et al. (2019)	**83.1**	81.8	81.1	81.1	81.5	82.2	81.1
	RACE-Middle	Lai et al. (2017)	74.1	73.8	74.1	**74.4**	73.3	73.3	72.7
	RACE-High	Lai et al. (2017)	69.4	69.3	**69.9**	69.7	69.2	68.8	69.7
	SciQ	Welbl et al. (2017)	94.0	95.5	95.0	98.0	96.6	94.1	**98.7**
	SocialIQA	Sap et al. (2019)	63.4	63.5	63.6	**64.2**	63.7	63.9	63.2
	BoolQ	Clark et al. (2019)	81.9	**82.2**	81.7	82.0	80.6	81.5	81.7
	MultiRC	Khashabi et al. (2018)	80.0	79.7	79.5	**80.0**	79.5	79.2	79.0
	WikiHop	Welbl et al. (2018)	58.7	58.7	58.8	**59.4**	59.4	59.1	58.5
	WIQA	Tandon et al. (2019)	74.4	74.4	75.1	**83.9**	83.6	82.4	83.5
Sentiment	IMDB	Maas et al. (2011)	94.8	**94.9**	94.7	94.9	94.7	94.7	94.8
	Rotten Tomatoes	Pang & Lee (2005)	90.2	89.6	**90.3**	89.9	90.0	90.0	89.6
Completion	HellaSwag	Zellers et al. (2019)	49.8	49.3	50.6	51.8	52.0	49.8	**53.7**
	COPA	Roemmele et al. (2011)	58.0	58.5	59.8	54.5	58.9	56.2	**62.0**
Topic Class.	AG News	Del Corso et al. (2005)	93.9	93.6	94.0	**94.3**	94.0	94.1	94.1
	DBpedia14	Lehmann et al. (2015)	**28.4**	28.4	28.4	28.4	28.4	28.4	28.4
WSD	WiC	Pilehvar et al. (2019)	68.8	67.9	69.5	**70.2**	68.2	66.3	69.8

- 実験設定
  - データセットごとに各知識源を利用
  - それぞれでfine-tuning, 知識なしと比較
- 結果：赤色が性能低下, 緑色が性能向上
  - タスクによって有効な知識源が異なる

  ⇒ **適切な知識源を選択する重要性**

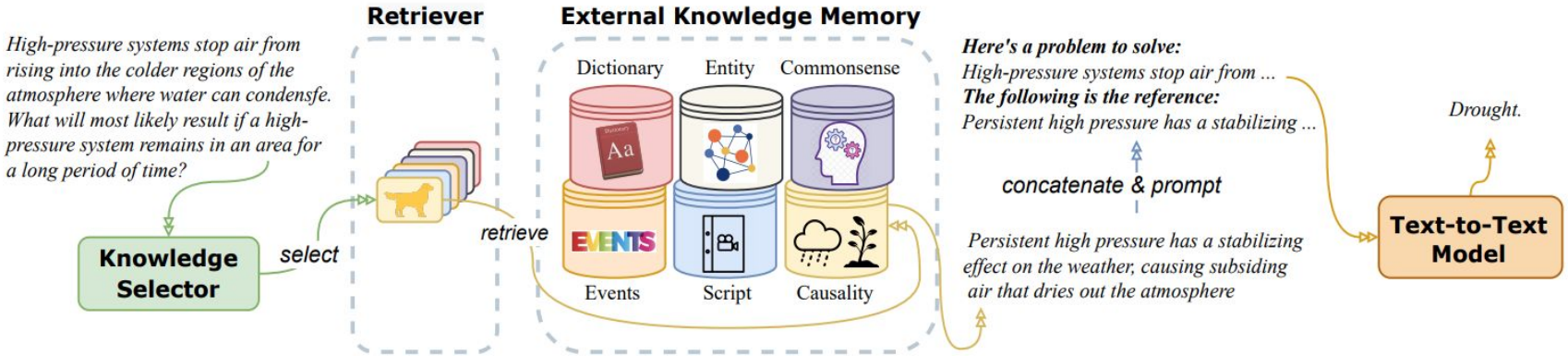
目標：タスクより細かい, インスタンスごとに

適切な知識源を選択する手法

Knowledge-in-Context（KiC）の全体像

- **6種類の構造化, 半構造化された外部知識源**を利用した言語モデル
  - Knowledge Selector：各入力インスタンスごとに適切な知識源の選択
  - Retriever：選択した知識源から入力に関連する知識を検索
  - Text-to-Text Model：検索した関連知識 + 入力文から回答を生成

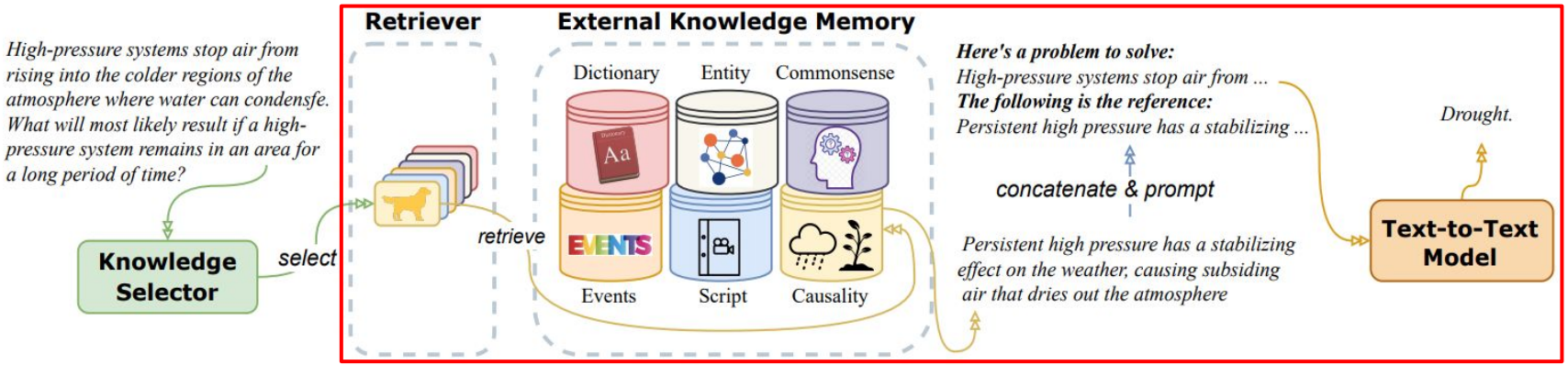
既存の方法を利用



Knowledge-in-Context（KiC）の全体像

- 6種類の構造化，半構造化された外部知識源を利用した言語モデル
  - Knowledge Selector：各入力インスタンスごとに適切な知識源の選択
  - **Retriever：選択した知識源から入力に関連する知識を検索**
  - **Text-to-Text Model：検索した関連知識 + 入力文から回答を生成**

既存の方法を利用



知識メモリの作成

- 知識メモリ：6種類の知識源のKeyのベクトルと対応するValue（自然言語）
  - Key-Valueのペアは，知識源ごとにそれぞれ定義
  - Keyはsentence encoderで密なベクトルに変換

各知識源の説明

知識源	説明
Dictionary	用語の定義と例文
Commonsense	一般的な知識，常識
Entity	エンティティに関する知識
Event	イベントに関する知識
Script	発話と身体の動き，声のトーン，表情などの非言語情報
Causality	因果関係に関する知識

例えば，Dictionaryでは，Key=用語，Value=定義

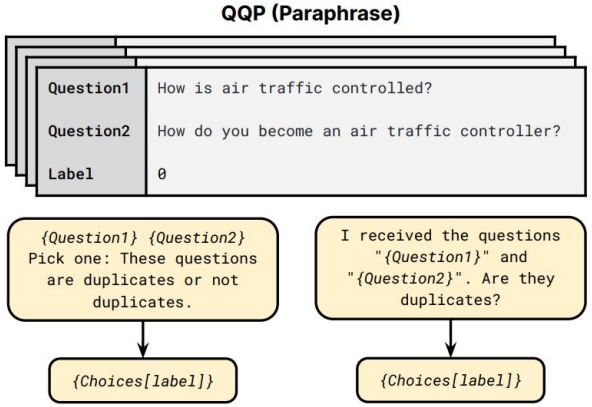
	Key	Value
Dictionary	s	o
Commonsense	s $s \oplus o$ $s \oplus r \oplus o$	$s \oplus r \oplus o$ $s \oplus r \oplus o$ $s \oplus r \oplus o$
Entity	s o	o o
Event	s $s \oplus o$ $s \oplus r \oplus o$	$s \oplus r \oplus o$ $s \oplus r \oplus o$ $s \oplus r \oplus o$
Script	s o	r r
Causality	$s \oplus o$ $o \oplus s$	$s \oplus o$ $o \oplus s$

Retriever

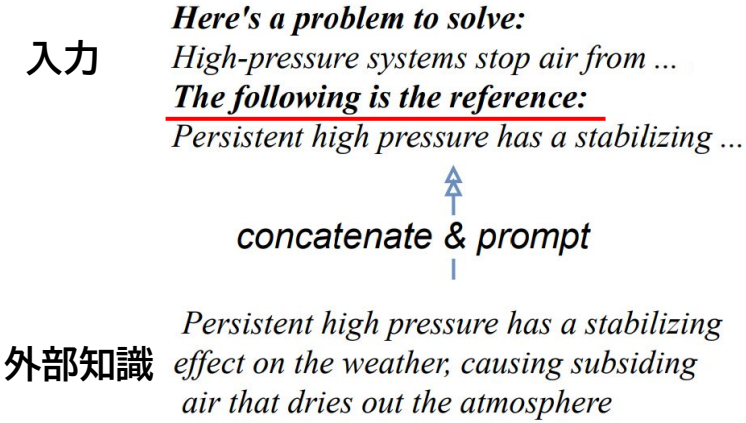
- Keyと同じSentence encoderで入力文をベクトル化
- 入力文のベクトルと内積が最大となるKeyを検索
- 一部の外部知識源の検索時には前処理
  - Dictionary：入力文から抽出したキーワードに関連する知識のみを検索対象
  - Entity：入力文から抽出したEntityとその関連するEntityのみを検索対象

Text-to-Text Model | プロンプトの作成

- 既存のテンプレートを使用
  - P3：各データセットに対して複数のテンプレート
  - 検索された外部知識（Value）はconcatするだけ？特に工夫なし



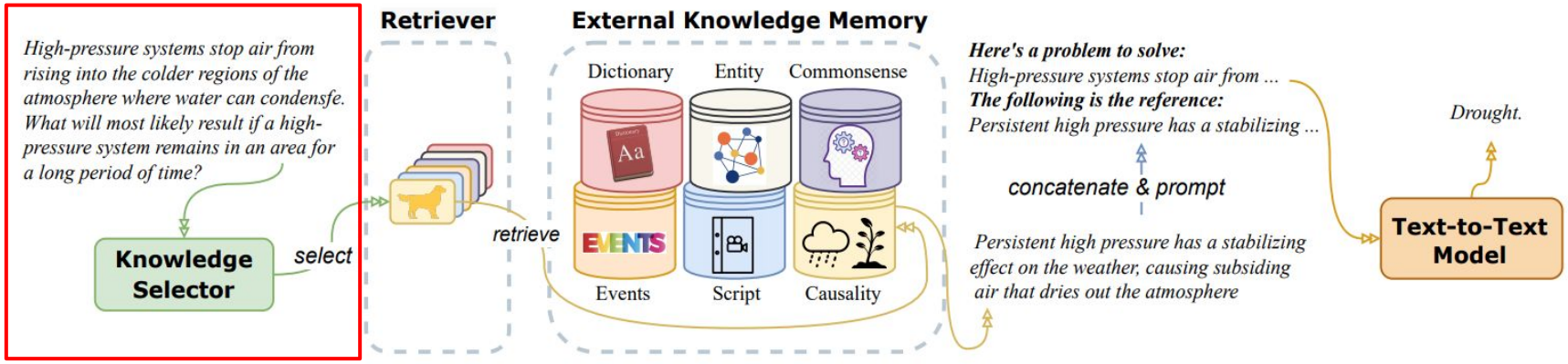
[Sanh et al., Multitask Prompted Training Enables Zero-Shot Task Generalization, ICLR2022]



Knowledge-in-Context（KiC）の全体像

- 6種類の構造化，半構造化された外部知識源を利用した言語モデル
  - **Knowledge Selector：各入力インスタンスごとに適切な知識源の選択**
  - Retriever：選択した知識源から入力に関連する知識を検索
  - Text-to-Text Model：検索した関連知識 + 入力文から回答を生成

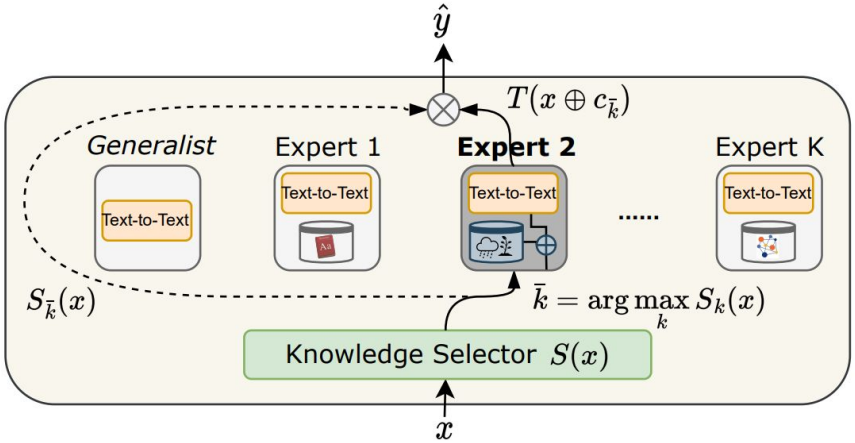
既存の方法を利用

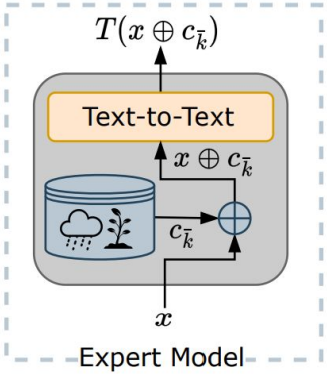


Knowledge Selectorの概要

- 各入力文ごとに適切な知識源の選択をする線形分類器
  - 入力文と適切な知識源の正解ラベルはないため, 単純には学習不可

    ⇒ **各知識源 + 言語モデルがExpert**のmixture-of-experts（MoE）とみなす
  - Knowledge Selectorはsequence-to-expertのマッピングをするRouter





一般的なMoE
各Expertが**異なるパラメータ**を持つ

本論文の手法:
共通の言語モデル + **異なる知識源**
non-parametric

⇒ **Expertごとに固有のパラメータなし**

Knowledge Selectorの学習

- Switch Transformerなどの既存のMoE学習手法を利用

x：入力文，S(x)：各知識源を選択する確率，
c_k：知識源kから検索した外部知識（**自然言語**）

離散値

予測

$$\bar{k} = \arg\max_k S_k(x) \qquad \hat{y} = T(x \oplus c_{\bar{k}})$$

損失

$$\mathcal{L}(x, y) = \sum_{t=1}^{T} \texttt{CrossEntropy}\big(\hat{y}_t, y_t\big) + \alpha \cdot \texttt{Balancing}\big(S(x)\big)$$

多様な選択をするようにBalancing lossを追加

Balancing loss

$$\texttt{Balancing}\big(S(x)\big) = (K+1) \cdot \sum_{i=0}^{K+1} f_i \cdot P_i, \qquad f_i = \frac{1}{B} \sum_{x \in \mathcal{B}} \mathbb{1}\Big(\arg\max_k S_k(x) = i\Big), \quad P_i = \frac{1}{B} \sum_{x \in \mathcal{B}} S_i(x).$$

Knowledge Selectorの学習

- Switch Transformerなどの既存のMoE学習手法を利用

x：入力文，S(x)：各知識源を選択する確率，
c_k：知識源kから検索した外部知識（**自然言語**）

離散値

予測

$$\bar{k} = \arg\max_k S_k(x) \qquad \hat{y} = T(x \oplus c_{\bar{k}}) \cdot S_{\bar{k}}(x)$$

$S_k(x)$をかけることで学習可能

損失

$$\mathcal{L}(x, y) = \sum_{t=1}^{T} \texttt{CrossEntropy}(\hat{y}_t, y_t) + \alpha \cdot \texttt{Balancing}(S(x))$$

多様な選択をするようにBalancing lossを追加

Balancing loss

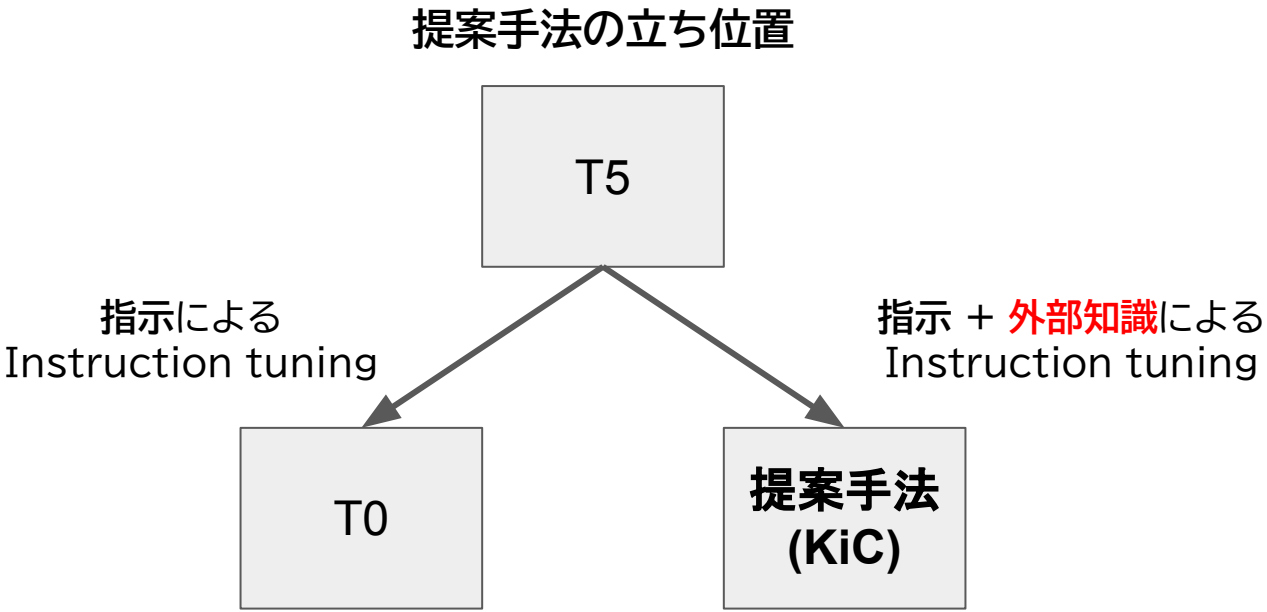
$$\texttt{Balancing}(S(x)) = (K+1) \cdot \sum_{i=0}^{K+1} f_i \cdot P_i, \qquad f_i = \frac{1}{B} \sum_{x \in \mathcal{B}} \mathbb{1}\Big(\arg\max_k S_k(x) = i\Big), \quad P_i = \frac{1}{B} \sum_{x \in \mathcal{B}} S_i(x).$$

実験設定

- モデル
  - Key, 入力文を埋め込むエンコーダー：All-MPNet$_{\text{base-v2}}$
  - 内積が最大となるKeyを検索する手法： SCaNN
  - Text-to-Text Model：T5$_{\text{LM-adapt}}$
- 学習データ：P3に含まれる39のタスク（840万件）
- 評価：Instruction tuningの学習に含まれない未知のタスクでのzero-shot
  - P3（Coreference resolution, NLI, Sentence completion, WSD）
  - MMLU（言語モデルが持つ知識を評価するデータセット）

実験設定

- 比較手法:T0, GPT3などの言語モデル



実験結果

- $\text{KiC}_{\text{Large}}$(0.77B)が3B$\text{T0}_{\text{3B}}$を凌駕
  - $\text{KiC}_{\text{Large}}$と$\text{T0}_{\text{Large}}$は同じモデル構造で，外部知識のあり/なし
- 175BのGPT-3と比較してもそれほど悪くない．．．か？

$\text{KiC}_{\text{Large}}$はGPT-3から
4.5ポイント減

P3での結果

Models	Params	Coreference WSC	Coreference Wino.	NLI ANLI_{R1}	NLI ANLI_{R2}	NLI ANLI_{R3}	NLI CB	NLI RTE	Completion COPA	Completion H.S.	Completion S.C.	WSD WiC
T0_{Base}	0.22B	$61.1_{5.5}$	$50.6_{0.9}$	$32.2_{1.4}$	$33.0_{1.0}$	$34.2_{0.7}$	$53.6_{17.1}$	$64.1_{1.4}$	$65.7_{5.7}$	$25.7_{0.5}$	$80.6_{1.3}$	$50.7_{1.4}$
T0_{Large}	0.77B	$59.1_{5.9}$	$50.5_{0.3}$	$30.5_{2.0}$	$32.7_{0.6}$	$33.8_{0.8}$	$60.7_{23.0}$	$62.1_{3.1}$	$73.5_{8.4}$	$25.7_{0.4}$	$84.1_{1.8}$	$50.2_{1.0}$
T0_{3B}	3B	$64.4_{2.7}$	$50.5_{1.2}$	$33.7_{0.9}$	$33.4_{1.2}$	$33.3_{0.4}$	$50.0_{15.9}$	$64.1_{3.5}$	$74.9_{8.7}$	$27.5_{1.0}$	$85.1_{3.2}$	$50.4_{0.9}$
T0_{11B}	11B	$64.4_{6.3}$	$60.5_{2.5}$	$44.7_{3.6}$	$39.4_{2.2}$	$42.4_{3.0}$	$78.6_{18.5}$	$81.2_{3.7}$	$90.8_{4.1}$	$33.7_{0.5}$	$94.7_{4.7}$	$57.2_{1.8}$
$\text{KiC}_{\text{Small}}$	0.06B	$63.5_{3.9}$	$51.1_{0.6}$	$33.3_{1.0}$	$33.3_{0.9}$	$33.6_{0.6}$	$44.6_{12.1}$	$47.3_{2.4}$	$48.0_{5.2}$	$25.4_{0.5}$	$57.7_{1.7}$	$50.0_{0.5}$
KiC_{Base}	0.22B	$63.5_{1.0}$	$50.0_{0.4}$	$28.4_{2.4}$	$30.9_{1.7}$	$32.8_{1.1}$	$58.9_{17.2}$	$66.8_{2.9}$	$65.0_{9.0}$	$26.1_{0.7}$	$82.6_{0.8}$	$50.2_{1.5}$
$\text{KiC}_{\text{Large}}$	0.77B	$65.4_{8.3}$	$55.3_{2.4}$	$36.3_{1.8}$	$35.0_{1.4}$	$37.6_{2.5}$	$67.9_{22.9}$	$74.0_{3.8}$	$85.3_{6.8}$	$29.6_{0.9}$	$94.4_{1.2}$	$52.4_{1.5}$

MMLUでの結果

Models	Params	Method	Average
$\text{RoBERTa}_{\text{Large}}$	0.35B	fine-tune	27.9
GPT-2	1.5B	fine-tune	32.4
Gopher	7.1B	5-shot	29.5
Atlas	11B	5-shot	47.9
GPT-3	13B	5-shot	26.0
GPT-NeoX	20B	5-shot	33.6
GPT-3	175B	5-shot	43.9
GPT-Neo	2.7B	0-shot	26.1
T0_{3B}	3B	0-shot	35.7
GPT-J	6B	0-shot	28.2
T0_{11B}	11B	0-shot	43.2
Atlas	11B	0-shot	47.5
GPT-NeoX	20B	0-shot	28.7
OPT	30B	0-shot	28.4
$\text{KiC}_{\text{Small}}$	0.06B	0-shot	26.8
KiC_{Base}	0.22B	0-shot	31.4
$\text{KiC}_{\text{Large}}$	0.77B	0-shot	39.4

Ablation Study

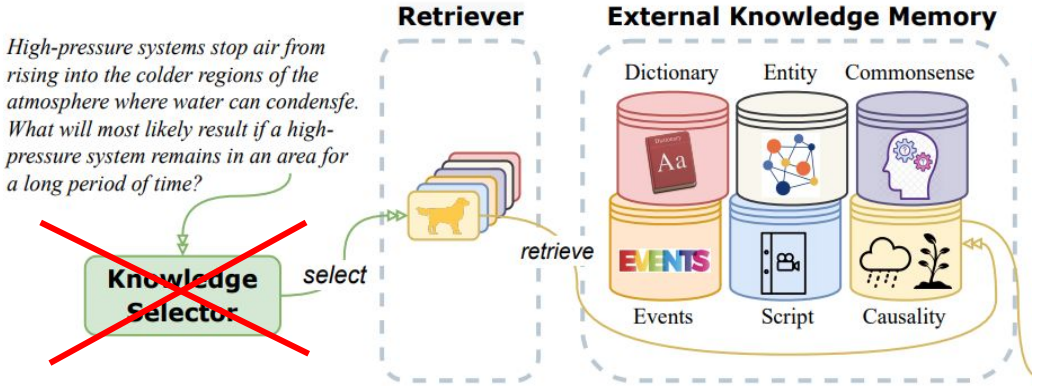
- 知識なし, テキストコーパス使用の場合と比べて性能向上

Dataset	Task	$\text{KiC}_{\text{Large}}$			w/o knowledge			/w plain texts		
		mean	med.	std	mean	med.	std	mean	med.	std
P3	WSC	62.6	65.4	8.3	57.6	59.1	5.9	53.3	52.9	7.7
	Wino. XL	54.1	55.3	2.4	50.4	50.5	0.3	52.2	52.2	0.4
	ANLI_{R1}	36.7	36.3	1.8	31.3	30.5	2.0	34.0	33.9	0.8
	ANLI_{R2}	34.9	35.0	1.4	32.8	32.7	0.6	33.7	33.9	0.8
	ANLI_{R3}	37.6	37.6	2.5	34.0	33.8	0.8	37.0	37.8	1.7
	CB	57.5	67.9	22.9	52.0	60.7	23.0	59.3	69.6	21.6
	RTE	73.1	74.0	3.8	62.4	62.1	3.1	67.6	67.1	2.7
	COPA	81.7	85.3	6.8	71.6	73.5	8.4	71.1	73.5	7.1
	HellaSwag	29.7	29.6	0.9	25.7	25.7	0.4	26.9	26.7	0.9
	StoryCloze	93.9	94.4	1.2	84.5	84.1	1.8	86.9	86.7	1.2
	WiC	52.1	52.4	1.5	50.2	50.2	1.0	51.6	51.5	0.6
	Average	55.8	57.6		50.2	51.2		52.1	53.3	
MMLU	STEM	30.7			28.2			29.1		
	Humanities	38.3			31.9			32		
	Soc. Sci.	43.6			33.2			36.6		
	Other	44.8			33.8			36.4		
	Average	39.4			31.8			33.5		

Ablation Study

- Knowledge Selectorなしのとき性能低下
  - Dictionary, Entityを検索するときの前処理ができないため

Dataset	Task	KiC_{Large}			w/o selector		
		mean	med.	std	mean	med.	std
P3	WSC	62.6	65.4	8.3	63.5	64.4	1.8
	Wino. XL	54.1	55.3	2.4	52.5	52.2	1.0
	$ANLI_{R1}$	36.7	36.3	1.8	31.6	31.5	1.1
	$ANLI_{R2}$	34.9	35.0	1.4	32.8	32.8	0.8
	$ANLI_{R3}$	37.6	37.6	2.5	33.9	34.0	1.2
	CB	57.5	67.9	22.9	52.7	62.5	20.3
	RTE	73.1	74.0	3.8	67.1	66.2	4.2
	COPA	81.7	85.3	6.8	75.7	79.0	7.3
	HellaSwag	29.7	29.6	0.9	28.4	28.4	0.4
	StoryCloze	93.9	94.4	1.2	90.4	91.1	1.6
	WiC	52.1	52.4	1.5	50.5	50.3	1.0
	Average	55.8	57.6		52.6	53.9	
MMLU	STEM	30.7			28.7		
	Humanities	38.3			36.3		
	Soc. Sci.	43.6			42		
	Other	44.8			42.7		
	Average	39.4			37.4		



Ablation Study

- タスクごとの知識選択よりもインスタンスごとの知識選択の方が性能が良い

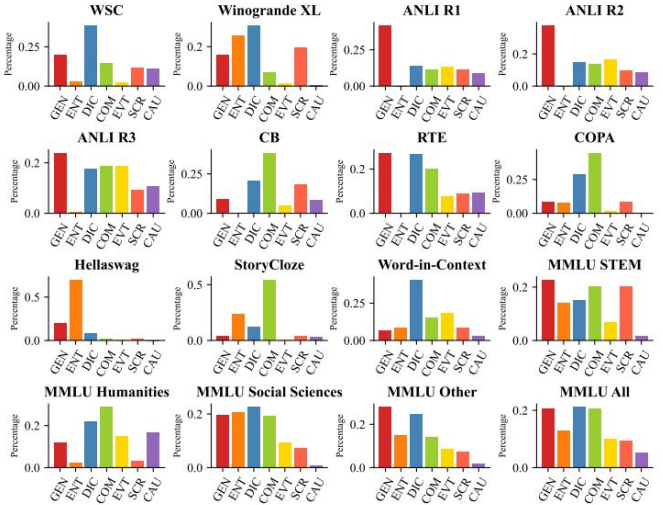
Dataset	Task	KiC_{Large}			/w task-adaptive		
		mean	med.	std	mean	med.	std
P3	WSC	62.6	65.4	8.3	62.2	63.9	5.0
	Wino. XL	54.1	55.3	2.4	53.8	54.7	2.2
	$ANLI_{R1}$	36.7	36.3	1.8	33.1	32.2	2.3
	$ANLI_{R2}$	34.9	35.0	1.4	33.7	33.1	1.8
	$ANLI_{R3}$	37.6	37.6	2.5	35.5	35.6	1.6
	CB	57.5	67.9	22.9	54.0	60.7	22.1
	RTE	73.1	74.0	3.8	73.1	73.3	4.1
	COPA	81.7	85.3	6.8	77.6	82.0	6.9
	HellaSwag	29.7	29.6	0.9	29.5	29.3	1.1
	StoryCloze	93.9	94.4	1.2	89.0	90.0	2.0
	WiC	52.1	52.4	1.5	52.1	51.2	2.4
	Average	55.8	57.6		54.0	55.1	
MMLU	STEM	30.7			29.2		
	Humanities	38.3			35.3		
	Soc. Sci.	43.6			40.5		
	Other	44.8			41.8		
	Average	39.4			36.7		

Ablation Study

- 外部知識を利用しないという選択肢(Generalist)も有効
  - 知識なしを選択する割合と比べて, 性能低下は小さい ⇒ 誤った知識は無視してる？

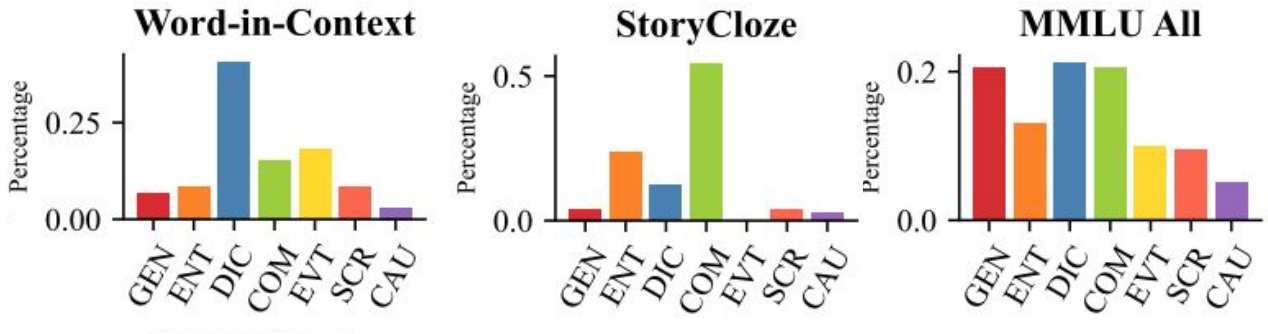
Dataset	Task	KiC_{Large}			w/o generalist		
		mean	med.	std	mean	med.	std
P3	WSC	62.6	65.4	8.3	62.8	64.4	7.5
	Wino. XL	54.1	55.3	2.4	54.2	54.9	2.0
	$ANLI_{R1}$	36.7	36.3	1.8	35.2	34.6	1.8
	$ANLI_{R2}$	34.9	35.0	1.4	35.1	34.9	1.1
	$ANLI_{R3}$	37.6	37.6	2.5	37.0	37.1	1.5
	CB	57.5	67.9	22.9	56.8	66.1	21.1
	RTE	73.1	74.0	3.8	73.0	72.4	2.7
	COPA	81.7	85.3	6.8	83.9	85.2	6.2
	HellaSwag	29.7	29.6	0.9	28.6	28.3	0.9
	StoryCloze	93.9	94.4	1.2	93.6	94.3	1.4
	WiC	52.1	52.4	1.5	52.0	50.8	2.4
	Average	55.8	57.6		55.6	56.6	
MMLU	STEM	30.7			30.3		
	Humanities	38.3			37.3		
	Soc. Sci.	43.6			42.5		
	Other	44.8			43.8		
	Average	39.4			38.5		

各データセットで選択した知識源の割合(**赤が知識なし**)



解析

- タスクごとの選択された知識の割合 ⇒ 直観に合う選択な気がする



知識選択の特徴

データセット	知識源の選択	選択の理由
Word-in-Context	**Dictionary**を最も多く選択	異なる語義の曖昧性解消ため
StoryCloze	**Commonsense**を最も多く選択	ストーリーの結末の完成のため
MMLU	多様な選択	広範な分野の質問に答えるため

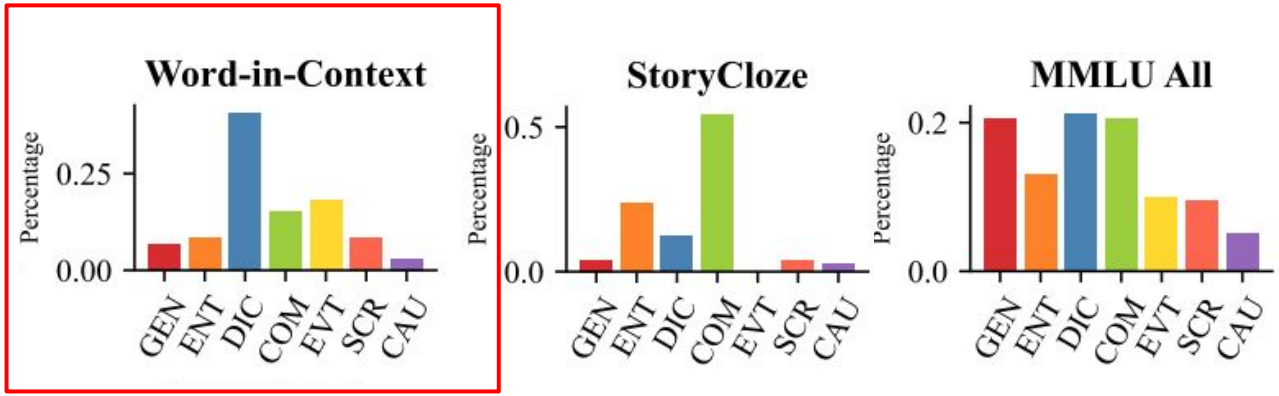
解析

（再掲）事前実験：異なる知識源の効果の違いを確認

Category	Task	Task Reference	None	ENT	DIC	COM	EVT	SCR	CAU
WSD	WiC	Pilehvar et al. (2019)	68.8	67.9	69.5	**70.2**	68.2	66.3	69.8

設定が違うので，比較は難しいが，
結構，傾向が違うような．．．？

- タスクごとの選択された知識の割合 ⇒ 直観に合う選択な気がする

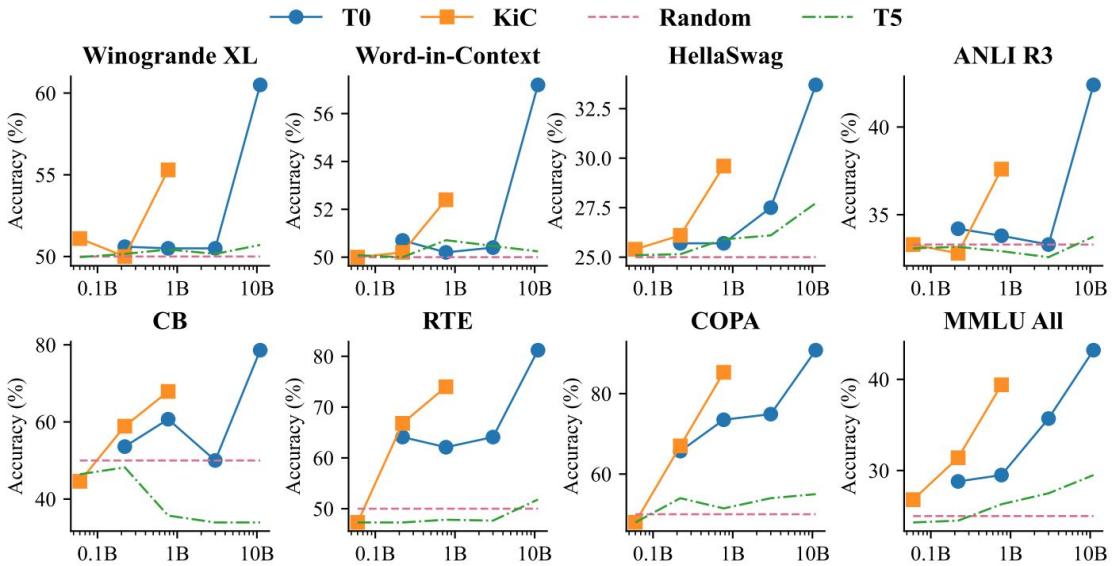


知識選択の特徴

データセット	知識源の選択	選択の理由
Word-in-Context	**Dictionary**を最も多く選択	異なる語義の曖昧性解消ため
StoryCloze	**Commonsense**を最も多く選択	ストーリーの結末の完成のため
MMLU	多様な選択	広範な分野の質問に答えるため

モデルサイズと性能

- 適切な外部知識源選択により，少ないパラメータでより高いzero-shot性能



T5：11Bまで大幅な性能向上が見られない
T0：3B ⇒ 11Bで大幅な性能向上
KiC：0.22B ⇒ 0.77Bで大幅な性能向上

まとめ

- 外部知識源から検索した知識を追加したInstruction tuning
  - 特徴1：**6種類の外部知識源から適切な知識源を選択**
  - 特徴2：**インスタンスごと**に適切な外部知識源を動的に選択
- 結果：テキストコーパスのみを利用する手法などと比べて性能向上

  ⇒ **インスタンスごとに適切な知識源を選択して利用する有用性**
- 今後の課題
  - 知識源の選択をTop 1からTop nへ
  - 知識源ごとのKey-Valueの定義など人手の作業がある程度必要

所感

- 感想
  - インスタンスごとに重要な知識源が異なることを確認したのは面白い
  - 複数の知識源からの外部知識を組み合わせて利用できると嬉しい
- 気になる点
  - 本当に各インスタンスごとに適切な知識源を選択できているのか？
    - もう少し解析があると嬉しい（知識源ごとの性能, 適切な選択をした時との比較など）
  - 学習中の選択の頻度はどうなっているのだろう？