# AI王 ～クイズAI日本一決定戦～ 第3回コンペティション

ICS Lab.（株式会社ベルシステム24ホールディングス）

2022/12/02

# ICS Lab.（株式会社ベルシステム24ホールディングス）について

**主な事業：**
コンタクトセンターアウトソーシング

**事業規模：**
- 国内39拠点，19,000席超の席数
- 31,000人以上のオペレーター
- 年間500,000,000コールを受ける

日々，“自然言語”で業務を遂行し，
“自然言語”の実践的な課題が生まれる会社

**株式会社ベルシステム24ホールディングス**

**沿革：**
2018～　Sony CSLとの共同研究開始
2020/4　ICS Lab.設立
2020/7　Mopas®, Knowledge Creator®提供開始

**設立趣意：**
現場での運用ノウハウ
x 機械学習・自然言語処理の実務適用
→「次世代コンタクトセンターの構築」

**ICS Lab.**
**（イノベーション＆コミュニケーションサイエンス研究所）**

# 問題設定の概要

- 問題設定
  - 質問の正解を文字列として出力する
    （第1回：正解を20の候補Entity=Wikipediaページから選ぶ）

**質問**

*映画『ウエスト・サイド物語』に登場する2つの少年グループといえば、シャーク団と何団?*

**解答**

ジェット団

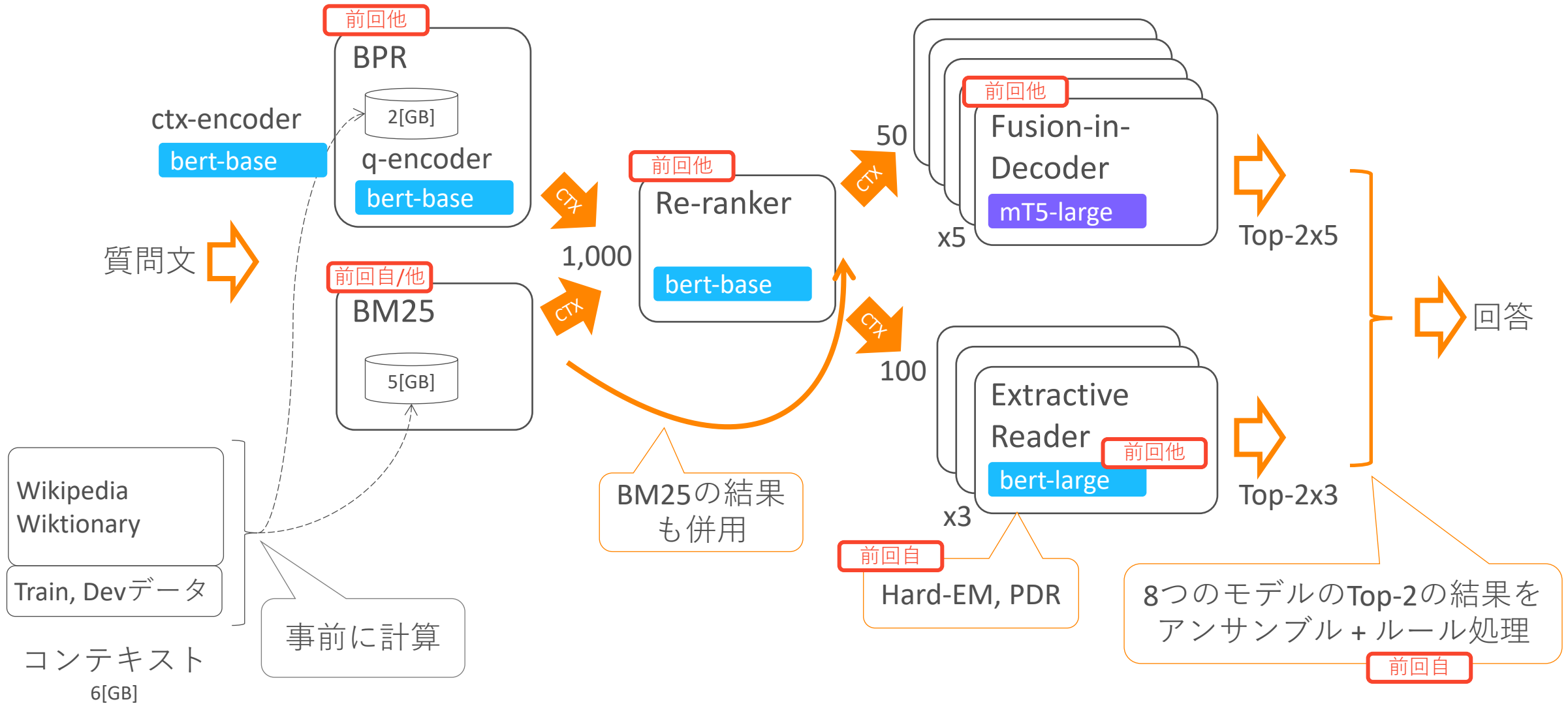
- その他のルール
  - 評価値 = 1,000問の正答率
    - 暫定評価：文字列の完全一致で確認
    - 最終評価：人手で表記の揺れなども考慮して確認
  - Wikipediaを含め，一般公開されている，もしくは公開できるデータのみ利用可能
  - 外部リソース（インターネット検索など）は利用禁止

# システム概要

前回自/他 : 前回{自, 他, 自/他}チーム同様の施策

mT5-large : google/mt5-large

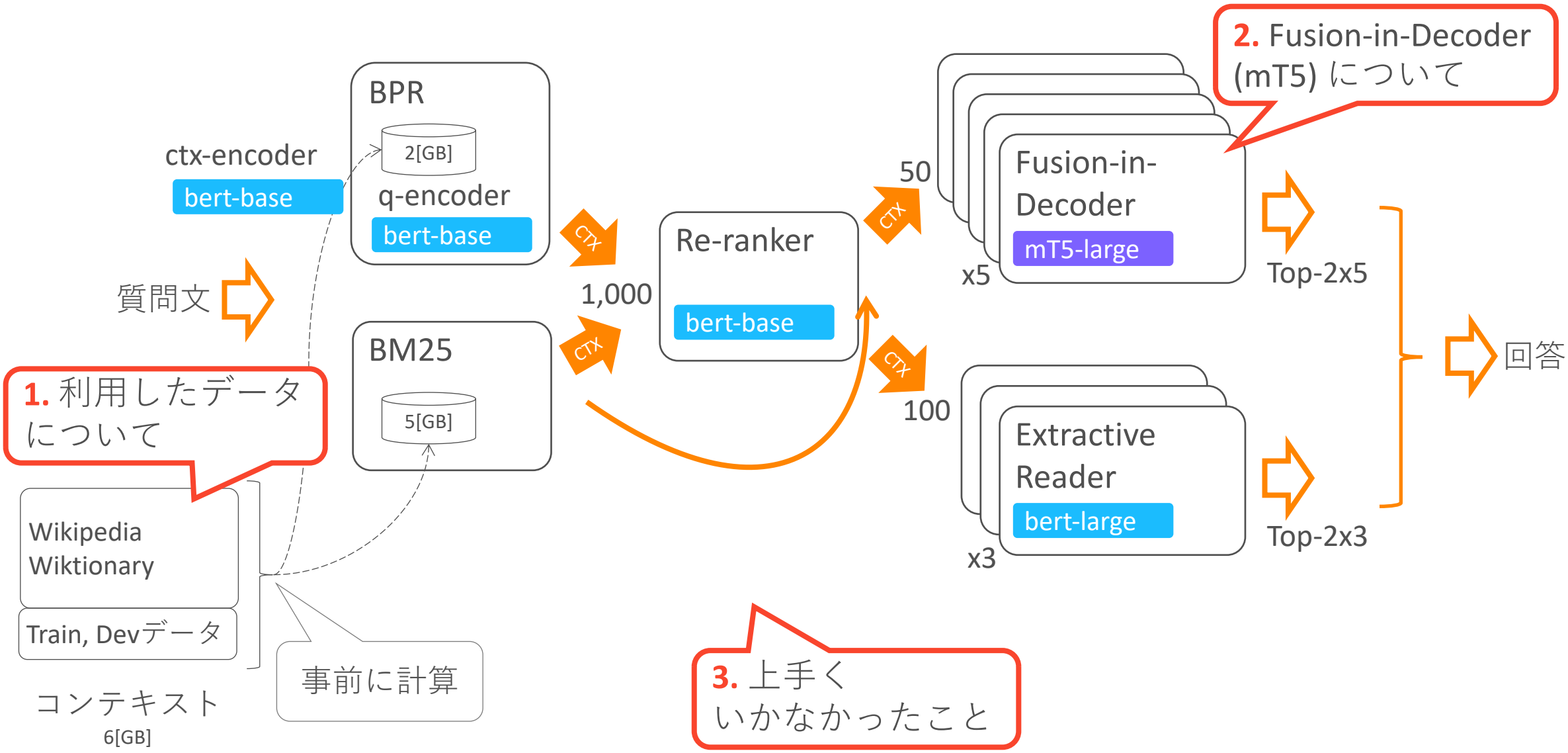
bert-?? : cl-tohoku/bert-{base, large}-Japanese(-v2)



# システム概要

mT5-large : google/mt5-large

bert-?? : cl-tohoku/bert-{base, large}-Japanese(-v2)



# 1. 利用したデータについて

前回自/他 : 前回{自, 他, 自/他}チーム同様の施策

- Q-A形式のデータをコンテキストにも利用（表の[*1]部分）
  - この形式を学習するため，10%を学習時にコンテキストに入れた

Q: 映画『ウエスト・サイド物語』に登場する2つの少年グループといえば、シャーク団と何団?
A: ジェット団

→

映画『ウエスト・サイド物語』に登場する2つの少年グループといえば、シャーク団と何団? 答えはジェット団。

| | | モデル学習時 | | アンサンブル重み決定時 | | 推論時 |
|---|---|---|---|---|---|---|
| | | コンテキスト | Q-Aデータ | コンテキスト | Q-Aデータ | コンテキスト |
| 公式trainデータ + クロールデータ[*2] | 90%分割 | | Yes | Yes[*1] | | Yes[*1] |
| | 10%分割 | Yes[*1] | | | Yes | Yes[*1] |
| 公式devデータ | | | Yes | Yes | | Yes |
| Wikipedia[*3] + Wiktionary | | Yes | | Yes | | Yes |

[*2] 前回他
Quiz Works: https://quiz-works.com/
みんはや: https://ss1.xrea.com/quizstocker.s1010.xrea.com/, https://livequiz.work/minhaya1/
語壷: http://www.misakichi.net/quiz/gogogo.htm
クイズの杜: https://quiz-schedule.info/quiz_no_mori/data/data.htm

[*3] 前回自
Wikipedia-Utils (https://github.com/singletongue/wikipedia-utils) を利用し，<p>だけでなく<dd><li>も利用

前回他

# 2. Fusion-in-Decoder(mT5)について

前回自/他 : 前回{自, 他, 自/他}チーム同様の施策

- mt5-largeの学習が大変
  - A100(メモリ40GBのGPU)x8を利用
  - DeepSpeed(stage=3), bf16, gradient checkpointingを利用
    - どれも省メモリで学習するための枠組み
    - GoogleのTPUで学習したモデルはfp16では上手く学習できず(A100ならbf16でできる)
  - 1回の学習にA100x8で，20~24時間ぐらい

- V100(メモリ16GBのGPU)での推論
  - 32bitでは，入力コンテキスト数=50，出力beam数=1が限界
  - fp16では上手く推論できなかったが，model.half()ならbeam数=2で推論できた
    - 出力beam数=1で，入力コンテキスト数を増やしたが精度は改善しなかった
    - 最終的にbeam数=2の結果を組み合わせてアンサンブルした

# 3. 上手くいかなかったこと

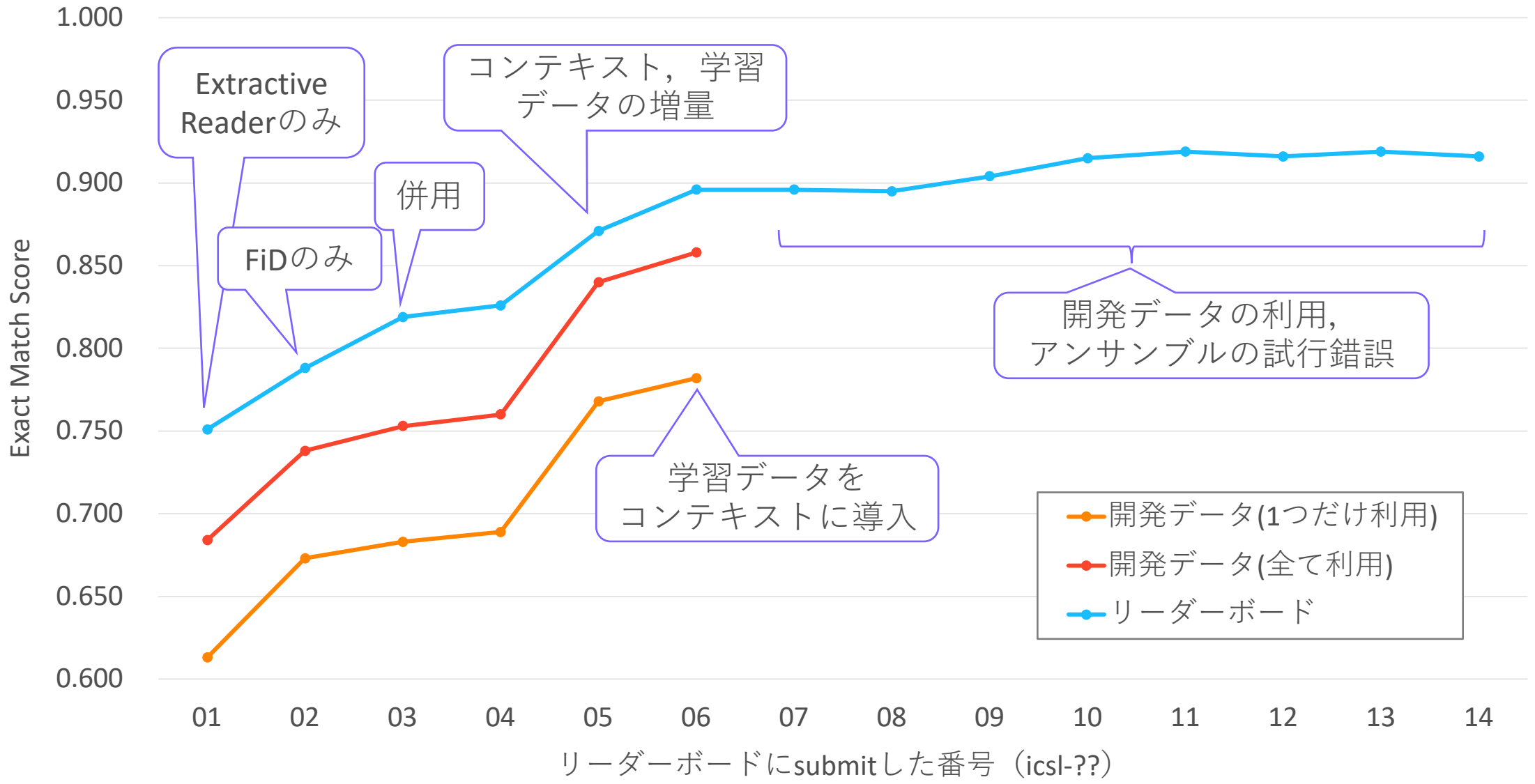
前回自/他 : 前回{自, 他, 自/他}チーム同様の施策

- FiD-KD: Izacard, Gautier, and Edouard Grave. "Distilling Knowledge from Reader to Retriever for Question Answering." *ICLR 2021, 9th International Conference on Learning Representations*. 2021.
  - Fusion-in-Decoderのattention scoreを見て，re-rankに利用
  - Scoreを取り出して，Answer含有率を確認したが良くなかった

- Hard Negative 前回自/他
  - Retriever(BPR)の学習・推論結果を使って再学習する手法
  - Retriever自体や，Extractive Readerは良くなるがFiDの精度が改善せず
- Re-rank結果を使ったReaderの学習
  - FiDの精度が改善せず

FiD学習時のコンテキストがPositive（答えが含まれるもの）が多すぎて
対照的な学習が出来なかった？
（もう少しNegativeな事例が入った方が良い？）

# スコアの推移（さまざまな施策が混ざっているためご参考まで）



# Future works

- FiD-KDとその先
  - 今回上手くいかなかったが，この延長（Readerの結果も使いながらRetrieverを一緒に学習する）が発展している

- より実践的？なタスク
  - Naturalな質問: 解答を思い浮かべない・知らない人が発する質問
  - Long-form QA: 質問・回答ともにある程度長い説明を必要とする質問
  - どちらもデータの作成，評価が難しいが．．．

> Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended.

ChatGPTのLimitationsより

- （引き続き）コールセンター業務への応用
  - Naturalで，Long-formで，質問に回答を特定するのに必要な情報が十分に含まれていない課題にどう応えていくのか？