

AI王 ～クイズAI日本一決定戦～ 第4回コンペティション 早押し解答部門，システム報告

ICS Lab. (株式会社ベルシステム24ホールディングス)
金本 勝吉，作本 猛

2024/01/26

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38拠点，19,000席超の席数
- 32,000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

- 1文字ずつ増えていく質問（クイズ）の答えを任意の文字列として出力する
- 正答するまで3回より多く誤答すると無得点（解答しないも可）
- スコア：正答すると +1点，早く答えると加算点
 - 加算点: 問題文の文字数(=N)に対してk文字目で正答すると $+(1.0 - k/N)$ 点

- その他のルール

- Wikipediaを含め，一般公開されている・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
 - Docker imageを提出し，外部通信できない状態で未知のクイズを解く
- 300問を12時間以内で解答し，最終評価は人手で採点

システム概要

mT5-large : google/mt5-large

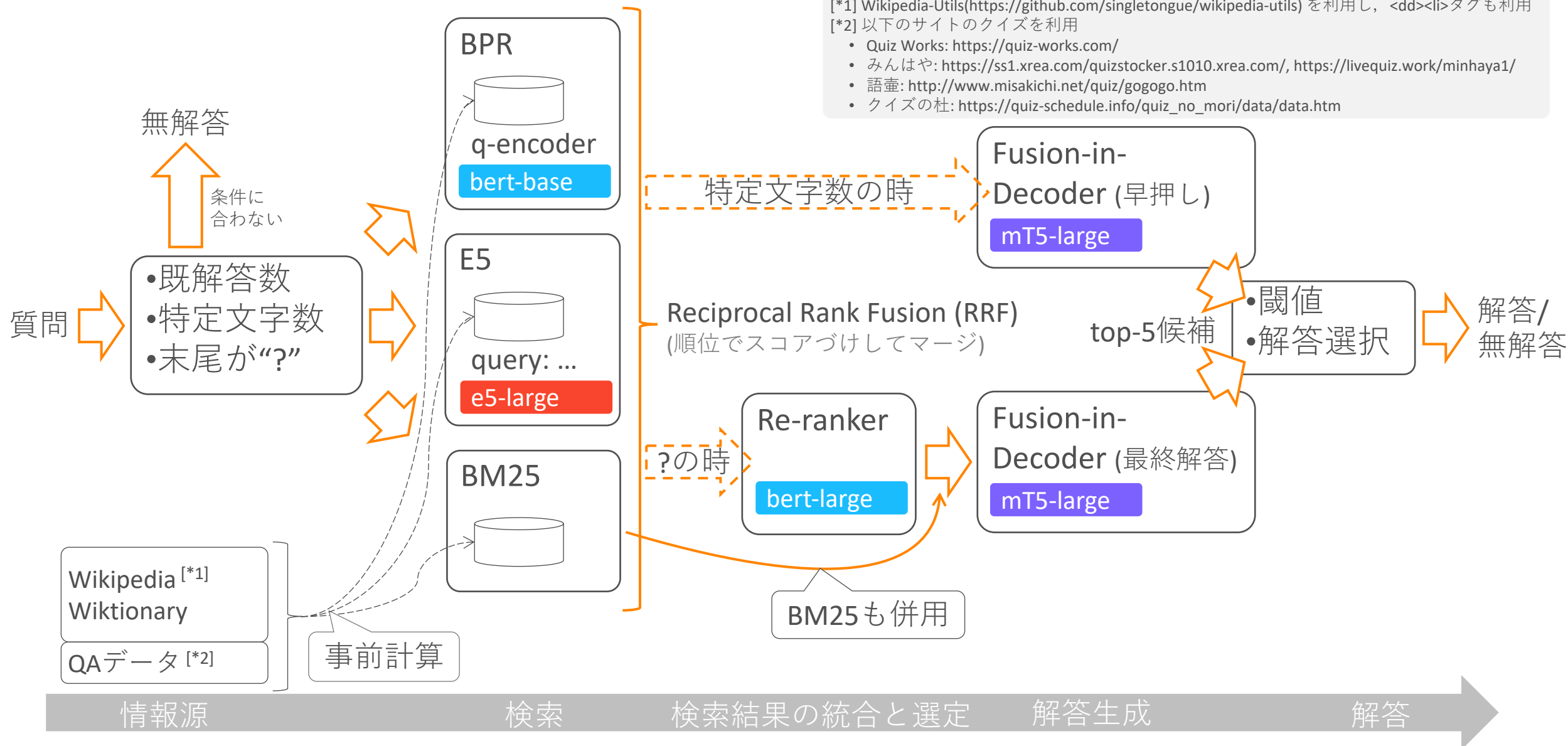
e5-large : intfloat/multilingual-e5-large

bert-?? : cl-tohoku/bert-{base, large}-Japanese(-v3)

[*1] Wikipedia-Utils(<https://github.com/singletonue/wikipedia-utils>) を利用し, <dd>タグも利用

[*2] 以下のサイトのクイズを利用

- Quiz Works: <https://quiz-works.com/>
- みんなはや: <https://ss1.xrea.com/quizstocker.s1010.xrea.com/>, <https://livequiz.work/minhaya1/>
- 語壺: <http://www.misakichi.net/quiz/gogogo.htm>
- クイズの杜: https://quiz-schedule.info/quiz_no_mori/data/data.htm

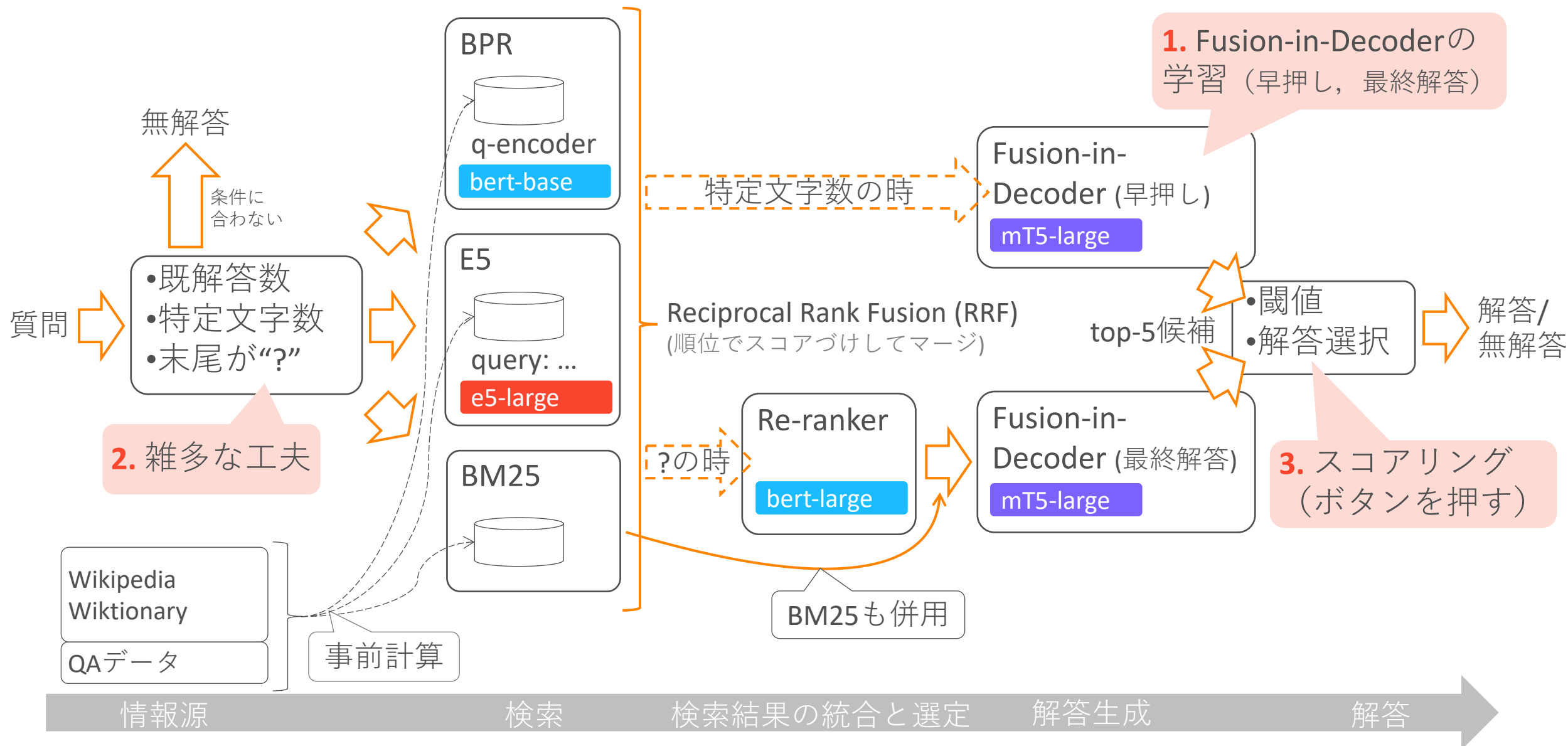


システム概要

mT5-large : google/mt5-large

e5-large : intfloat/multilingual-e5-large

bert-?? : cl-tohoku/bert-{base, large}-Japanese(-v3)



1. Fusion-in-Decoderの学習（早押し，最終解答）

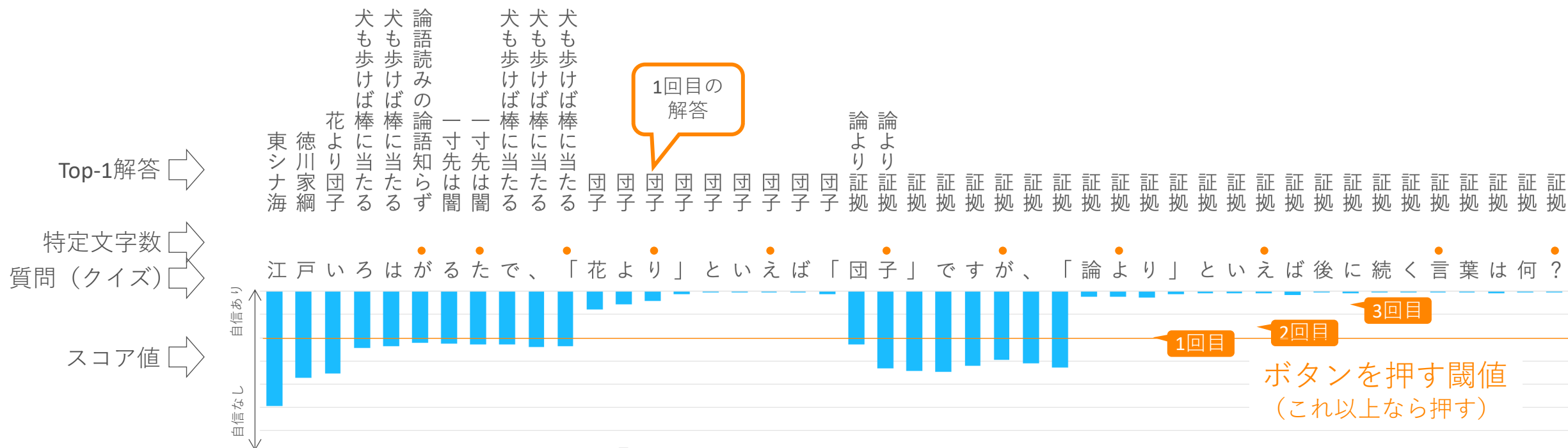
- Fusion-in-Decoderとは？
 - (質問+情報源)x50組を入力 → 解答となる文字列を生成するT5ベースのモデル
- 早押しモデルと最終解答モデルの2つを学習
 - 早押しモデル
 - 質問の途中でも解答を生成するモデル
 - 学習時にランダムに質問を削りながら正答を返すように学習
 - 学習時に与える情報源（コンテキスト）も，質問の途中で検索した結果を利用
 - 最終解答モデル
 - 質問が完全に与えられている前提で解答を生成するモデル
 - 90%超えで正答できる（第3回AI王の実績より）
 - “?”で終わっていればクイズの終端とみなしてこのモデルを利用（まれに例外あり）

2. 雑多な工夫

- 最大解答数の制限, 同じ解答の抑制
 - ルール上3回まで誤答してもペナルティなし → 解答は4回までに制御
 - 4回の解答でできるだけ違う解答を選択
 - Fusion-in-Decoderでtop-5で異なる解答を生成してから選択
- 特定文字数で解答 (何文字目で解答するか?)
 - 入力される1文字ずつ増える質問全てに解答すると間に合わない
 - 6, 8, 11, 14, 18, 22, 26, 30, 35, 41文字目と“?”の時のみ計算
- 最適化
 - 計算速度を上げるほど, 解答する特定文字数を増やせる → より早く解答できる
 - 最終的には早押し解答=12秒, 最終解答=20秒程度で解答 ($12 \times 10 + 20 = 140$ [秒/問])

3. スコアリング (ボタンを押す)

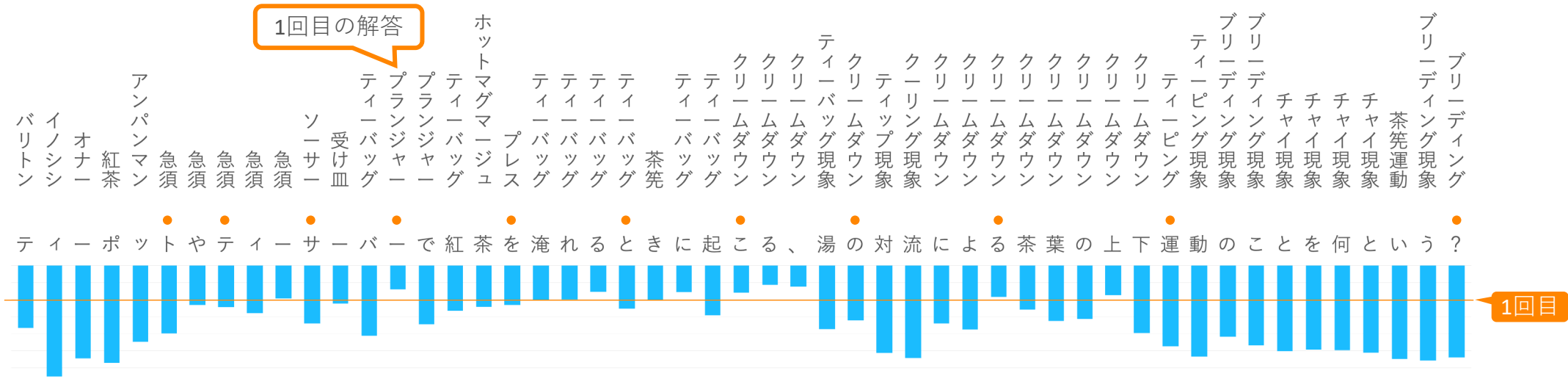
- 閾値でボタンを押すか（解答 or 無解答）を決定
 - 簡易的な計算で「解答可能性」を推定できないか検討したが困難だった
 - 結果的にFusion-in-Decoderの出力から算出するスコア値を利用
 - 1→2→3回目で徐々に閾値を上げた（＝最初は自信がなくてもボタンを押す）



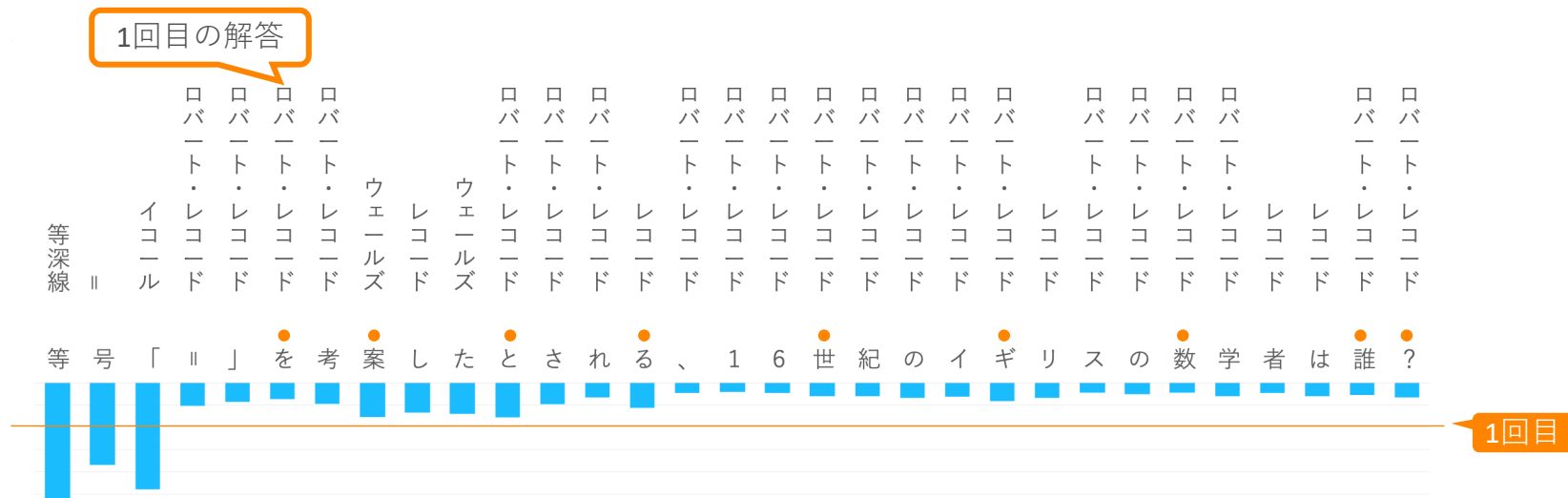
「ですが」問題におけるスコアの遷移例

(実際は計算時間の問題で特定文字数のみ計算)

3. スコアリング (ボタンを押す) 他 の 事 例

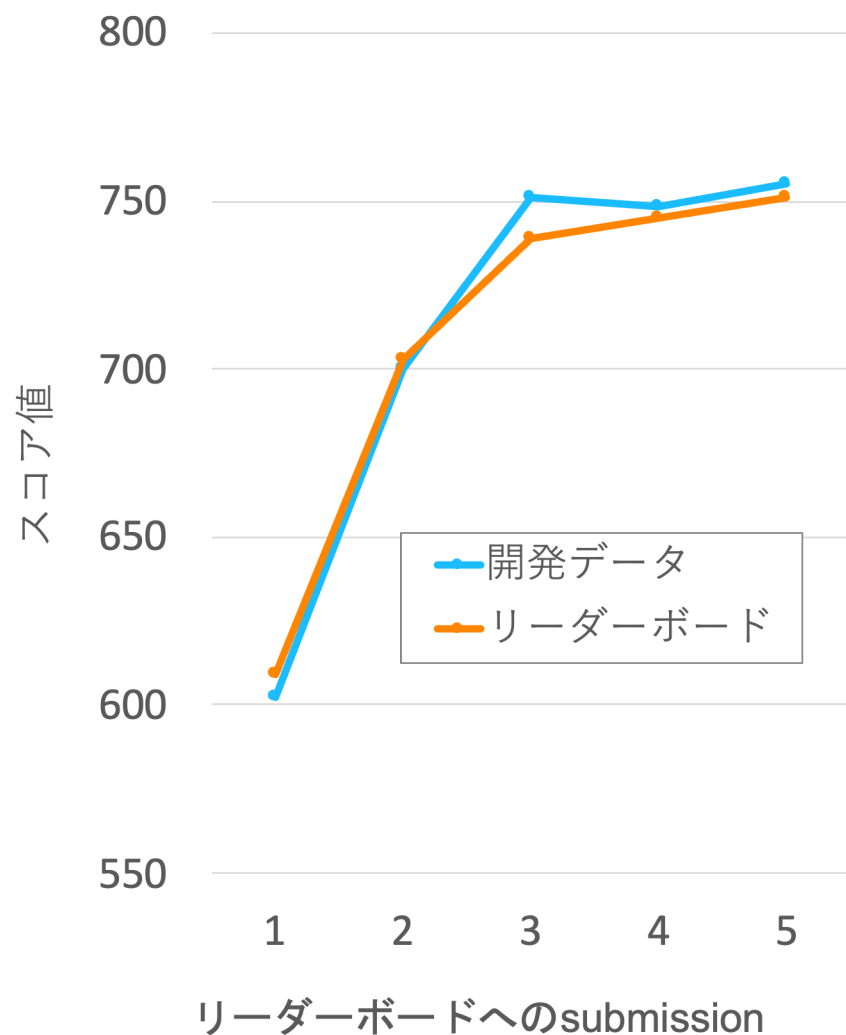


正答できなかった時のスコアの遷移例（正答＝ジャンピング）



早い段階で正答できた（はず）のスコアの遷移例（ウェールズは出身地）

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）



1	• 少なめのデータで学習したモデルの結果 • 固定文字数のみ解答（10, 20, 30, “?”のみ解答）
2	• データを増やして再学習 • E5の追加とRRFの導入（順位でスコアづけ） • 固定文字数のみ解答（10, 20, 30, “?”のみ解答）
3	• 特定文字数を6箇所+“?”に • 閾値を用いて早押し • Top-5候補から解答選択（重複を避ける）
4	• 特定文字数を6→8箇所+“?”に
5	• 特定文字数を8→10箇所+“?”に

おわりに

- まだまだ人間には及ばない...?
 - 3回間違えても良いルールだった． 誤答=0点 or ペナルティのルールだとどうか？
 - 1文字ごとに計算に12秒かかる． 計算資源をかければリアルタイムで解答できる？
- ChatGPTを含むLLMや汎用モデルでどこまで行ける？
 - GPT系のLLMで途中までの質問→質問全文の生成を試したがうまくいかず
 - 汎用モデルだと， 質問全文を与えられても9割も解答できない？
- （引き続き） コールセンター業務・実課題への応用
 - 早押しは， 必要な情報が十分に含まれていない質問でより実課題に近い
 - 経験豊富なオペレーターの方々は過去に聞かれた内容（＝経験）から足りない情報を補う
 - 一言で回答できない問題への対応（回答の生成）はLLMに期待？