AI王 ～クイズAI日本一決定戦～ 第2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）

2022/03/11

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37拠点，18,000席超の席数
- 30,000人以上のオペレーター
- 年間200,000,000コールを受けるクラウド基盤

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
    （前回：正解を20の候補Entity=Wikipediaページから選ぶ）

- その他のルール
  - 情報源，モデル等含めて圧縮済みで30[GB]以内
  - 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

質問

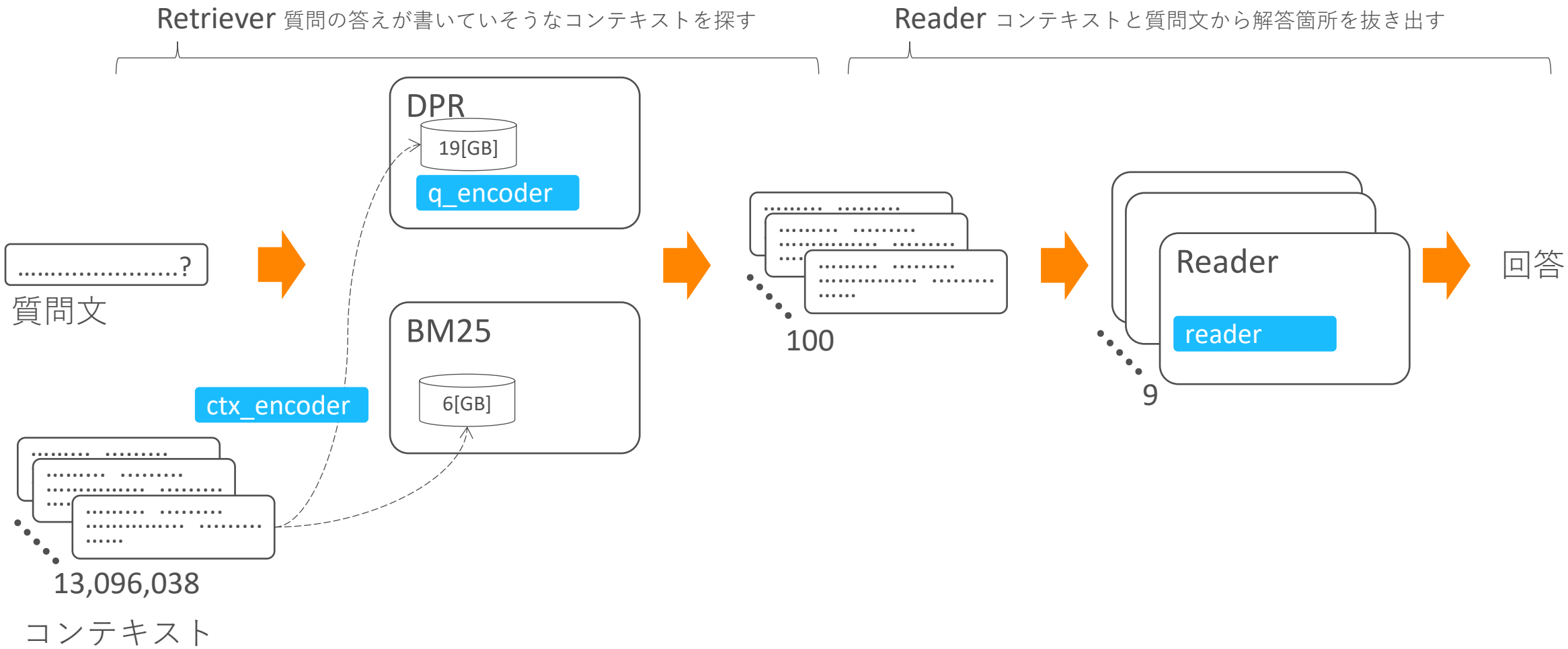
映画『ウエスト・サイド物語』に登場する2つの少年グループといえば、シャーク団と何団?

解答

ジェット団

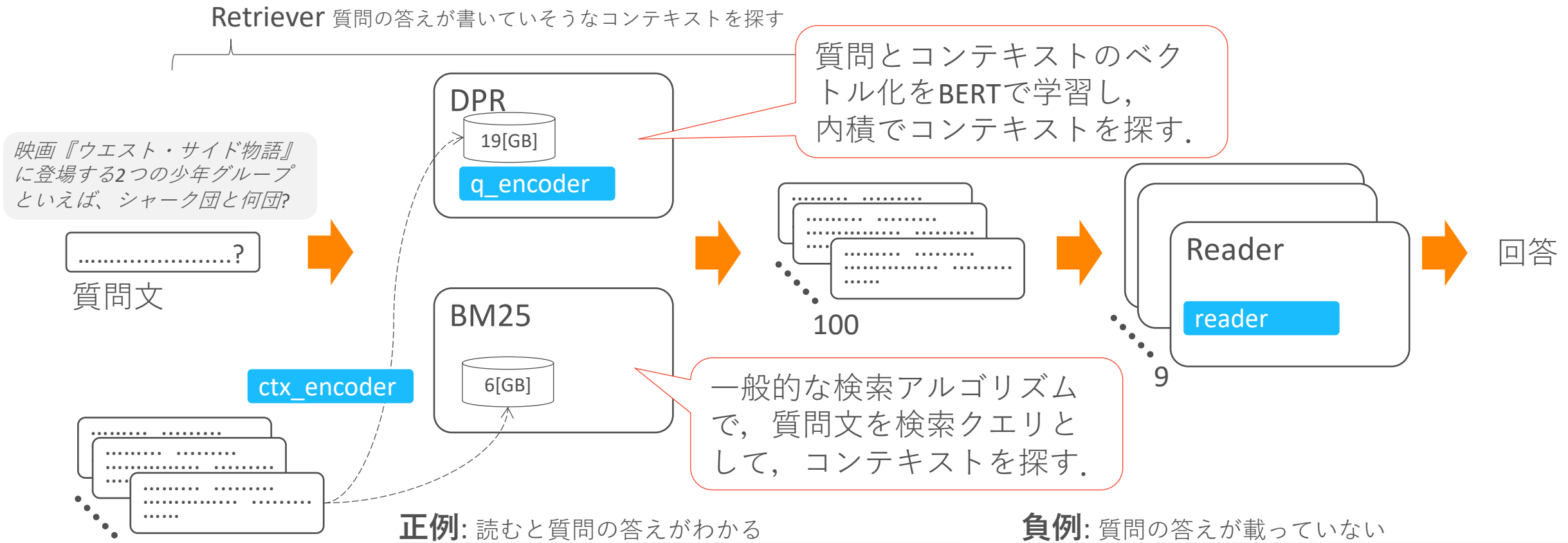
システム概要

: cl-tohoku/bert-base-japanese-v2ベースのBERTモデル



システム概要

: cl-tohoku/bert-base-japanese-v2ベースのBERTモデル



正例: 読むと質問の答えがわかる

ニューヨークのウエスト・サイド。午後5時。ポーランド系アメリカ人の少年非行グループ「ジェッツ」(ジェット団)と、新参のプエルトリコ系アメリカ人の少年非行グループ「シャークス」(シャーク団)は、なわばりを巡って対立している。

1957年の映画『青春物語(英語版)』で第30回アカデミー賞の助演男優賞にノミネートされる。1961年には『ウエスト・サイド物語』でジェット団のリーダー・リフ役を演じている。

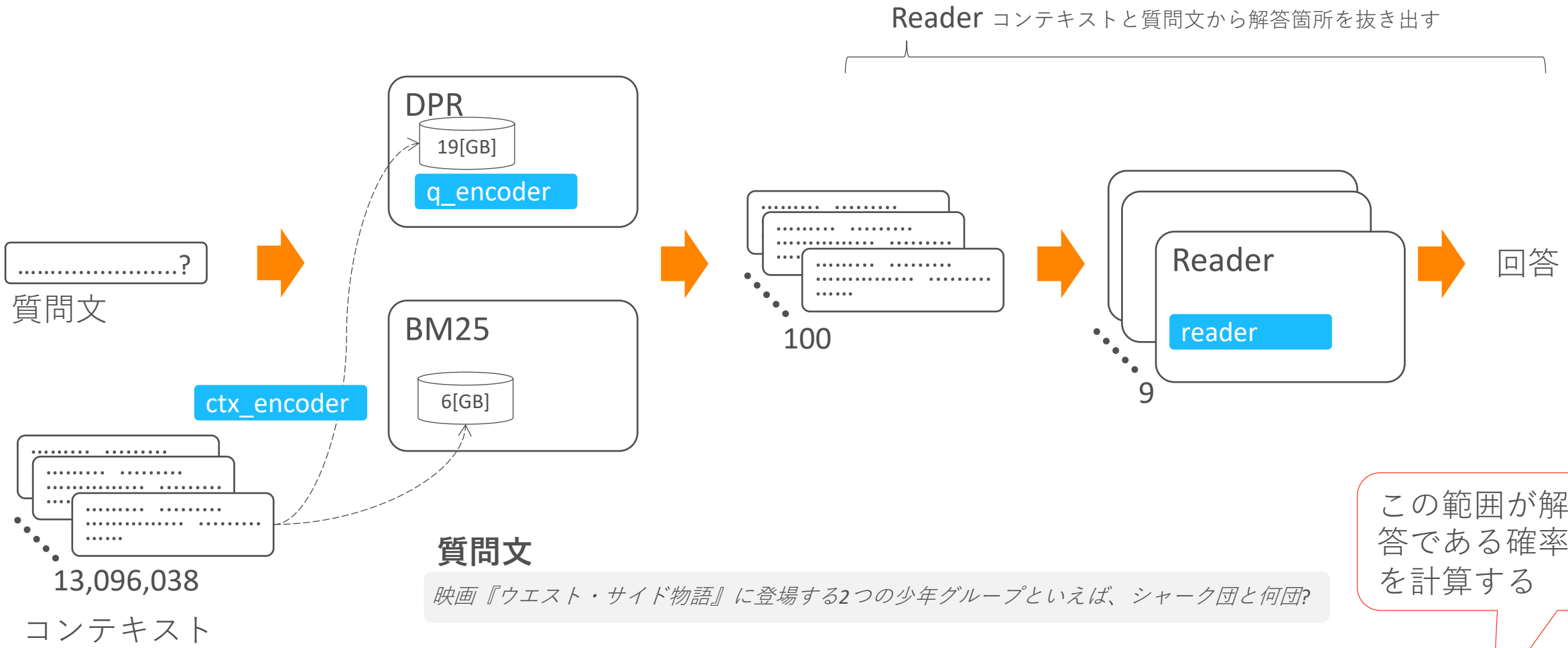
負例: 質問の答えが載っていない

ヒロイン・マリアの名前は、『ウエスト・サイド物語』(1961年)のシャーク団リーダー・ベルナルドの妹、および『サウンド・オブ・ミュージック』(1965年)の元修道女でトラップ家の家庭教師から取られている。

『ウエストサイド物語』(ウエストサイドものがたり)は、宝塚歌劇団によるミュージカル作品。ブロードウェイ・ミュージカルの傑作『ウエストサイド物語』の日本での上演の一つである。

システム概要

: cl-tohoku/bert-base-japanese-v2ベースのBERTモデル



質問文

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?

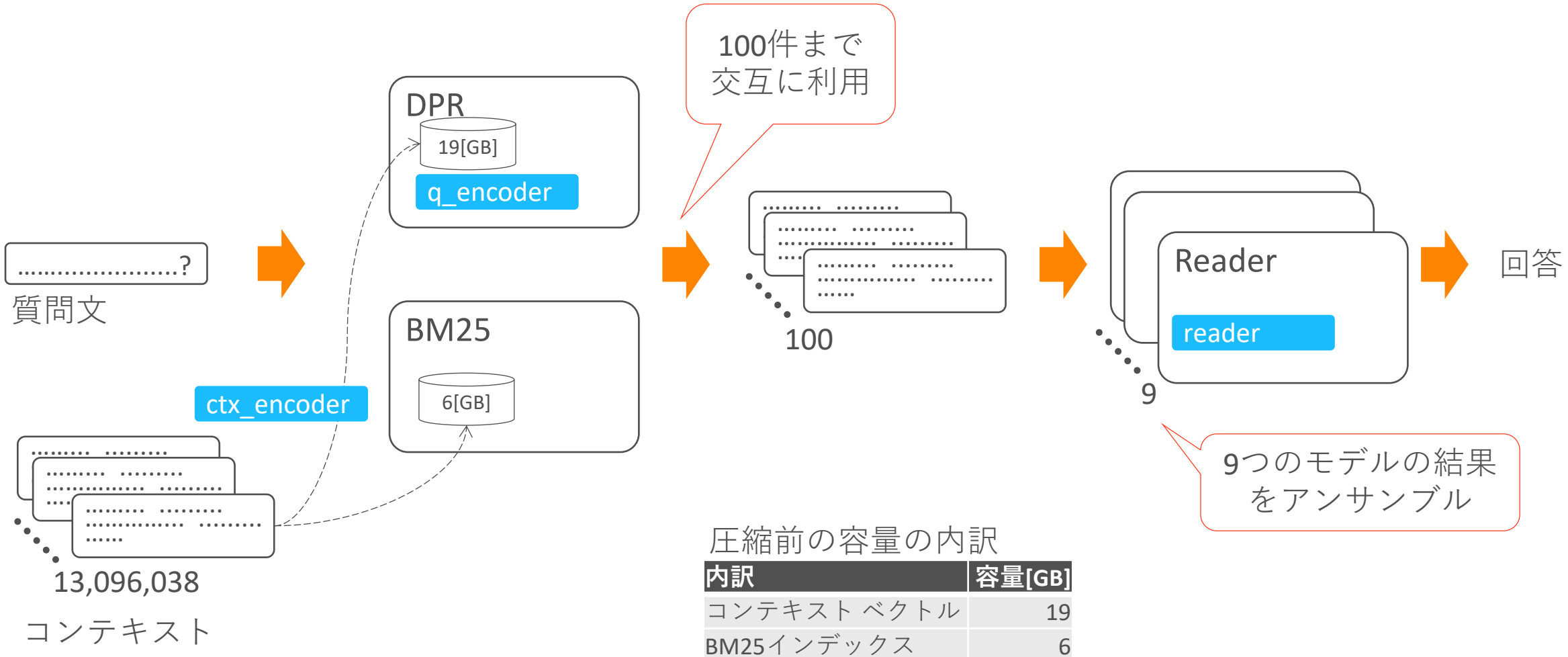
コンテキスト

ニューヨークのウエスト・サイド。午後5時。ポーランド系アメリカ人の少年非行グループ「ジェッツ」(ジェット団)と、新参のプエルトリコ系アメリカ人の少年非行グループ「シャークス」(シャーク団)は、なわばりを巡って対立している。

この範囲が解答である確率を計算する

システム概要

: cl-tohoku/bert-base-japanese-v2ベースのBERTモデル

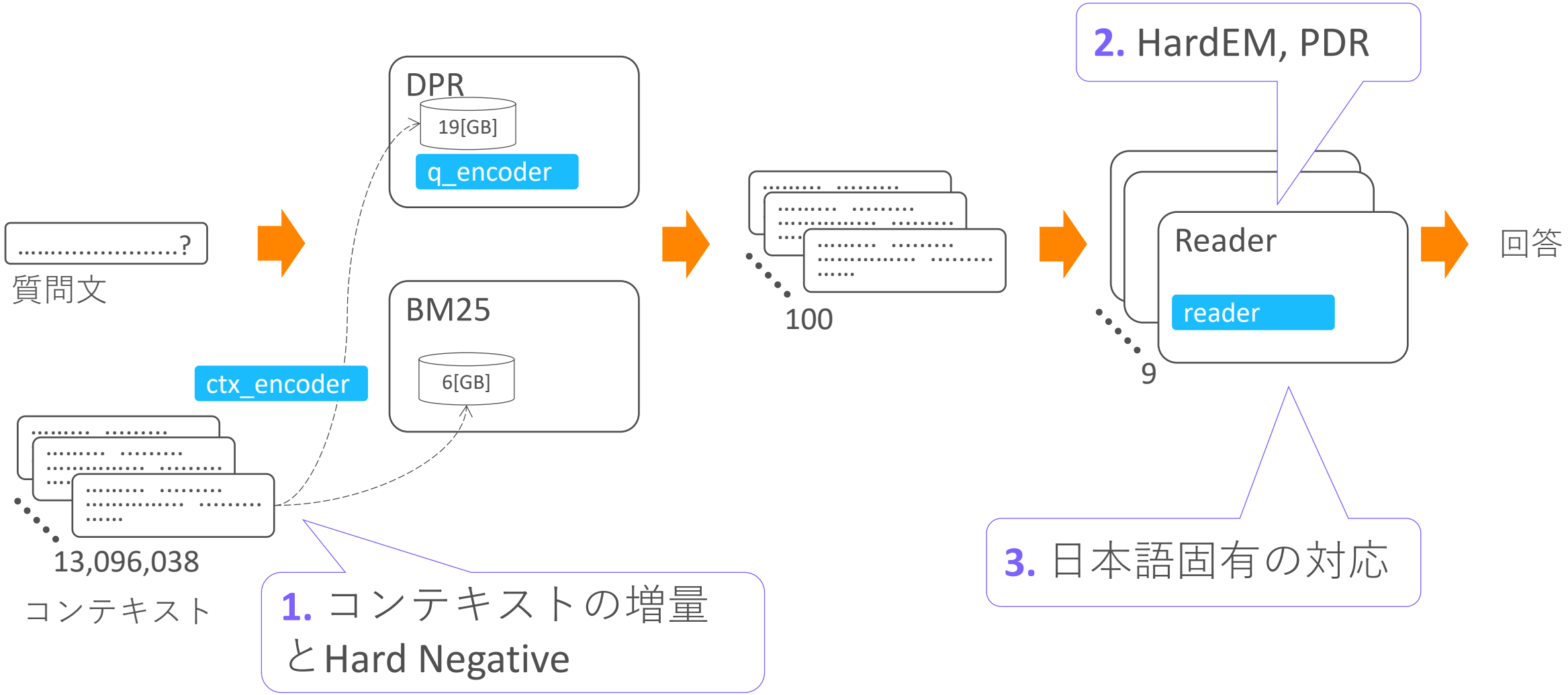


圧縮前の容量の内訳

内訳	容量[GB]
コンテキスト ベクトル	19
BM25インデックス	6
BERTモデル	4
OS,その他	3
コンテキスト 本文	2
合計	**34**

システム概要

: cl-tohoku/bert-base-japanese-v2ベースのBERTモデル

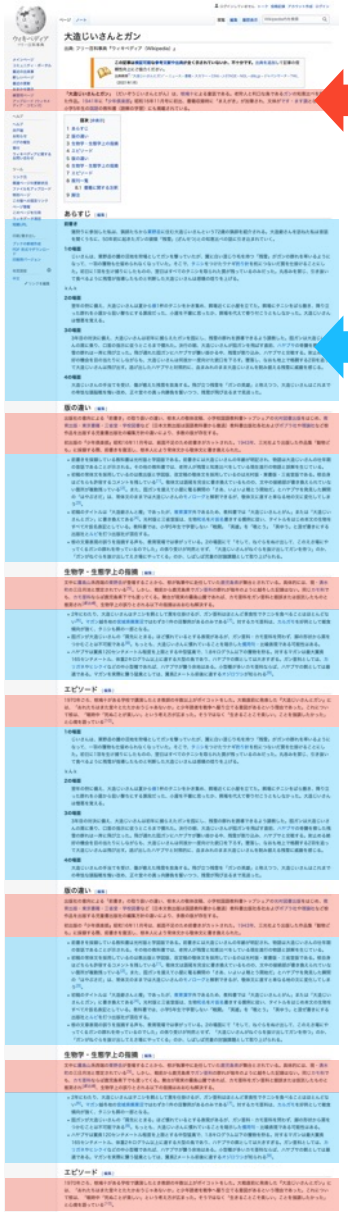


[*1] Masatoshi Suzuki , https://github.com/singletongue/wikipedia-utils

1. コンテキストの増量とHard Negative

- 「残雪」問題
  - 配布いただいたコンテキストは，<p>タグの内容のみ
    - 箇条書きなど，書き方によっては情報が入っていない
  - Wikipedia-Utils[*1]を利用し，<dd>
も利用

椋鳩十の童話『大造じいさんとガン』に登場する、ガンの群れの頭領の名前は何? → 残雪

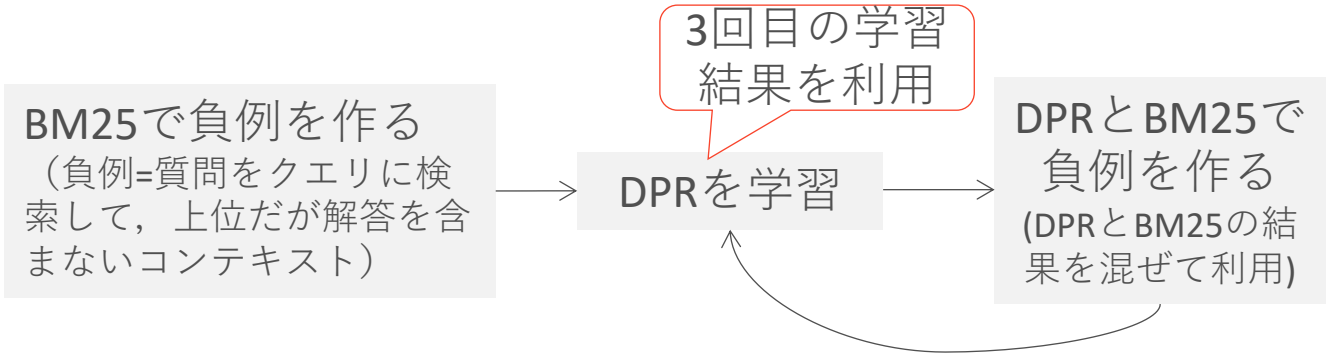


<p>で取れるコンテキスト

<dd>
で取れるコンテキスト

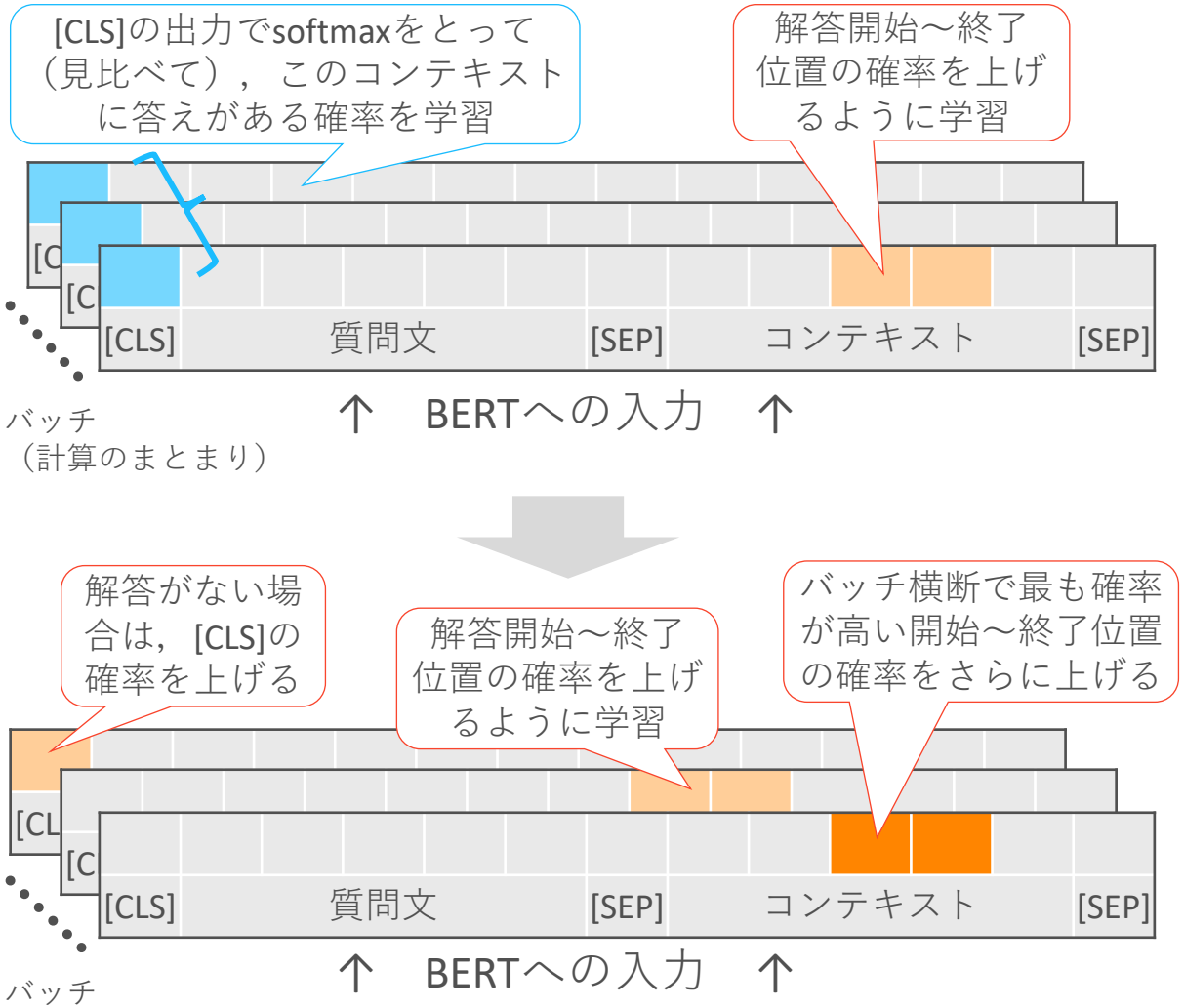
コンテキストの量が6百万
↓
13百万に倍増
↓
32→16bitで保存
ファイルサイズはほぼ同じ

- Hard Negative
  - DPR=正例と負例を見分けるタスク
    - 難しい負例で学習した方が見分けられるようになる
    - DPR自身の出力を使ってさらに難しい負例を作る

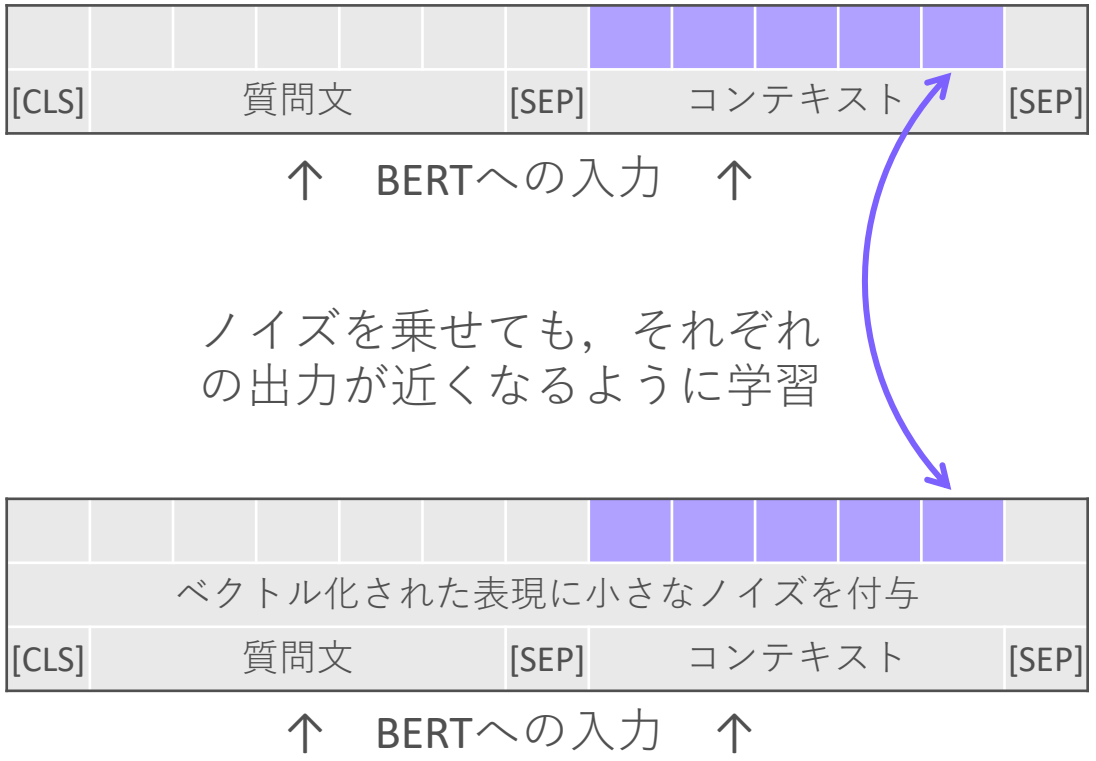


2. HardEM, PDR

[*2] Cheng, Hao, et al. "UnitedQA: A hybrid approach for open domain question answering." arXiv preprint arXiv:2101.00178 (2021).
[*3] Cheng, Hao, et al. "Probabilistic assumptions matter: Improved models for distantly-supervised document-level question answering." arXiv preprint arXiv:2005.01898 (2020).
[*4] Cheng, Hao, et al. "Posterior differential regularization with f-divergence for improving model robustness." arXiv preprint arXiv:2010.12638 (2020).



HardEM (Hard Exact Match) [*3][*2]



PDR (Posterior Differential Regularization) [*4][*2]

3. 日本語固有の対応（精度への影響は小さかった）

- Tokenizeのバリエーション
  - 同じ言葉が異なるID列にTokenizeされる
    ＝ 学習時に正解spanを特定できない

  →“##”がついている語彙，ついていない語彙
  両方の組み合わせのID列を生成してspanを特定

- [UNK]を読みで展開する
  - 難しい漢字は語彙に入っておらず，[UNK]（=不明なtoken）になってしまうが，読みに置き換えることで回避する

- n文字で何という？
  - Tokenize後に「文字数」の情報は全く入らない

  →正規表現（ルール）で候補から条件に合うものを選択

「王女アルテミスを」をtokenize

Token	王女	アルテ	ミス	を
ID列	18729	25025	13782	932

「アルテミス」をtokenize

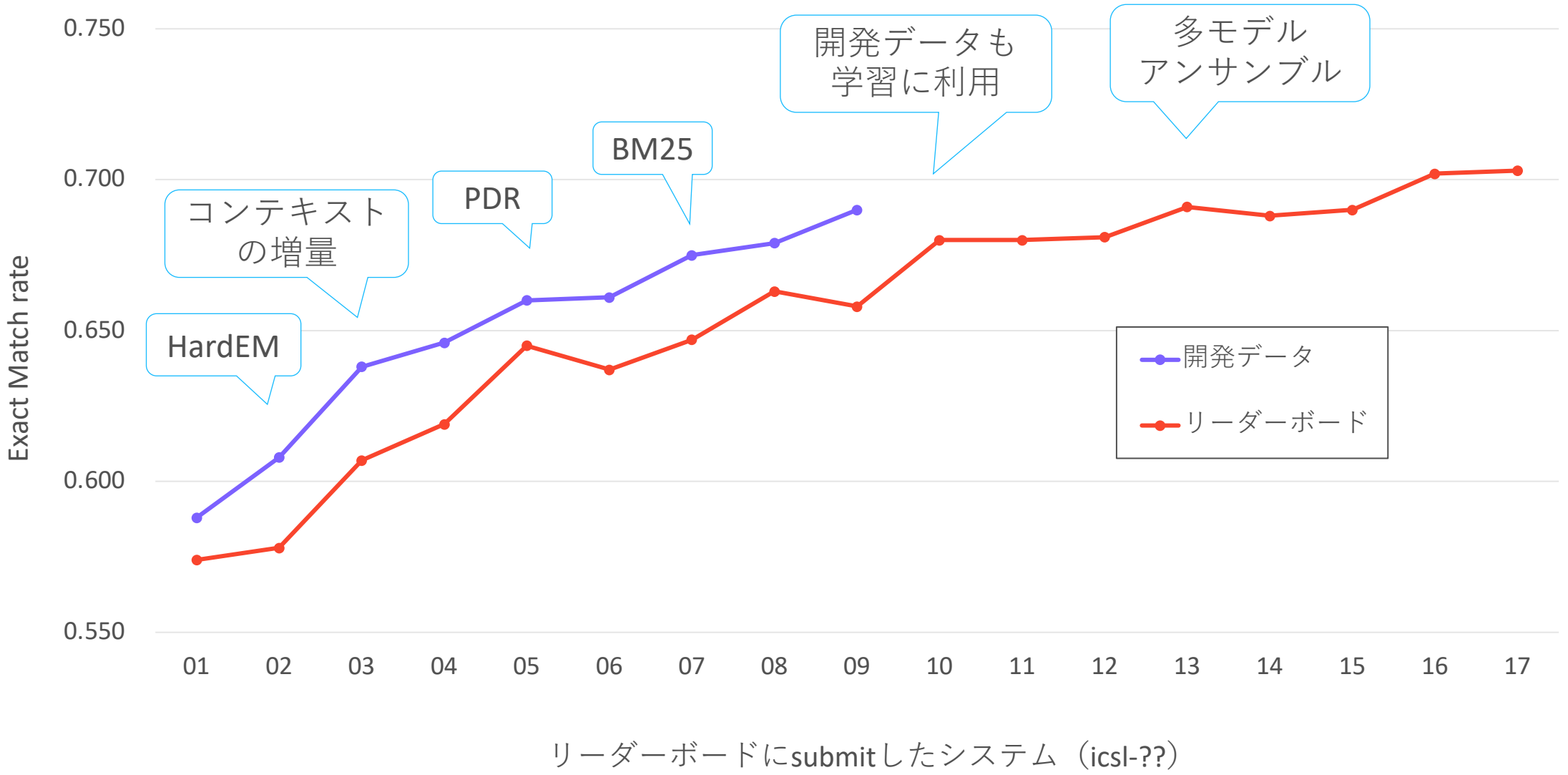
Token	アルテ	##ミス
ID列	25025	14879

「和同開珎」をtokenize

Token	和	##同	[UNK]
ID列	1604	8285	1

Token	和	##同	かい	##ちん
ID列	1604	8285	16392	32128

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 work

- コンテキストの更なる増量
  - ことわざ・慣用句などクイズによく出るが，Wikipediaに情報がない分野がある
- 生成型Readerモデル
  - 日本語のよく学習されたBARTやT5などのpre-trainedモデルがあれば一般知識をカバーできる可能性がある？
- コールセンター業務への応用
  - クイズはどちらかというと“Natural”ではない質問
    - 答えてほしい解答の特定に対し，そこに誘導するような問いや，比較的十分な情報がある
      - *…消化器官は何？／…選ばれる文学賞は何？／…ミュージカル映画は何？*
  - 情報が不十分で，一言で答えられない課題にどう応えていくのか？