レトリバ

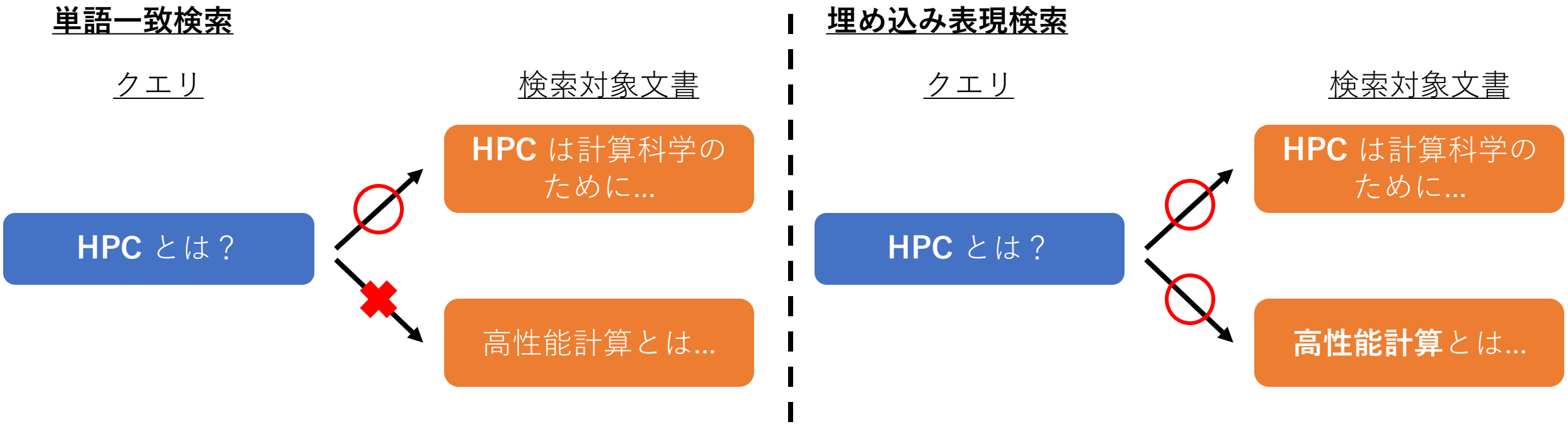
日本語向けの
テキスト埋め込みモデルの構築

株式会社レトリバ
勝又 智
西鳥羽 二郎

2025年6月11日

テキスト埋め込みモデルとは

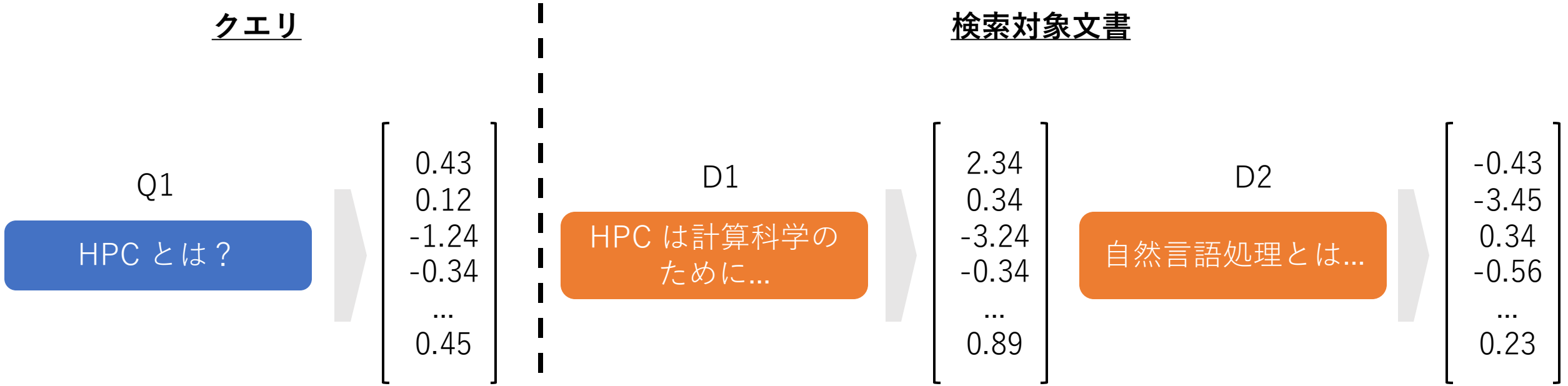
- 自然言語処理、情報検索の分野では「テキスト埋め込みモデル（Text Embedding）」の開発が盛んに行われている
- 応用先
  - 情報検索
  - RAG（Retrieval Augmented Generation）
- 従来の単語一致を利用した情報検索と違い、埋め込まれた表現を用いることで、意味的な類似度を扱うことが可能になる



テキスト埋め込みモデルの構築

レトリバ

- テキスト埋め込みモデルは、与えられたテキストに対して何らかのベクトル表現を作成する
- 本取り組みでは、「テキスト埋め込みモデルの構築」によって、与えられたテキストに対して適切な意味を持つ埋め込みを作成するモデルを作成することを目指す
- 以下の例では、「Q1 と D1 の類似度は Q1 と D2 の類似度より大きい」が理想的

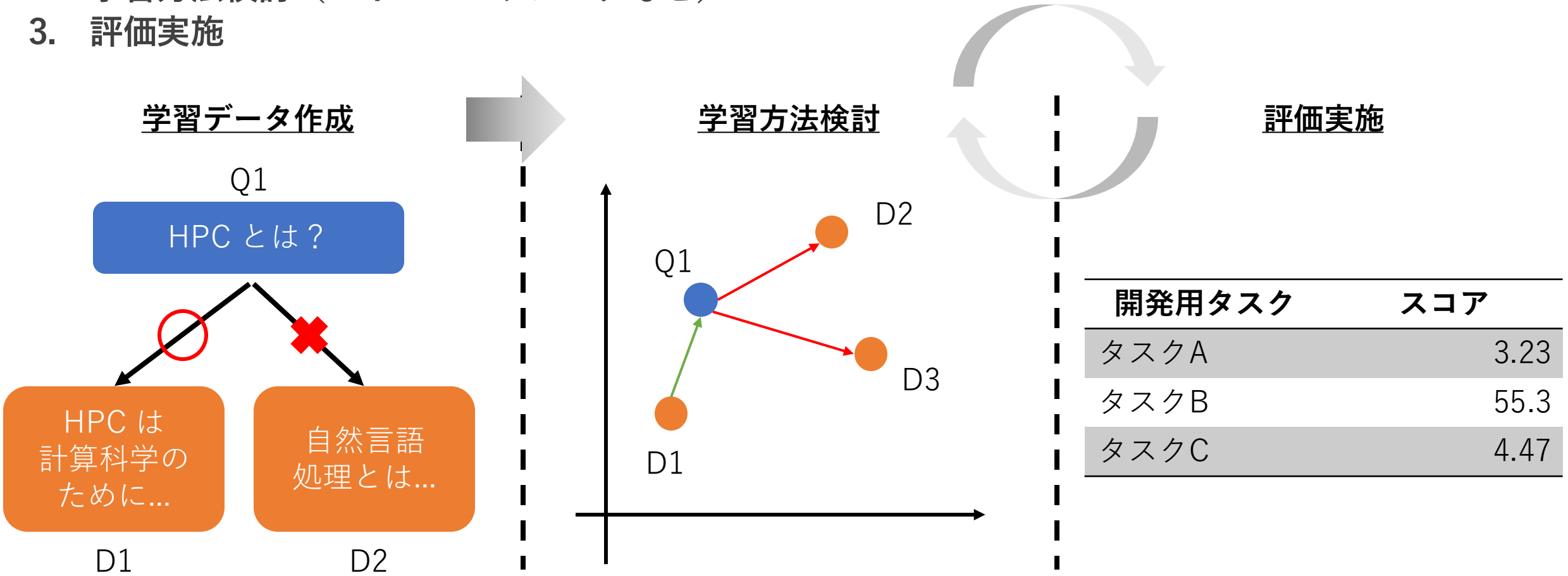


Similarity(Vec(Q1), Vec(D1)) > Similarity(Vec(Q1), Vec(D2))

テキスト埋め込みモデルの構築に向けた取り組み

テキスト埋め込みモデル構築に向けて、次の取り組みを実施した。

1. 学習データ作成
2. 学習方法検討（ハイパーパラメータなど）
3. 評価実施

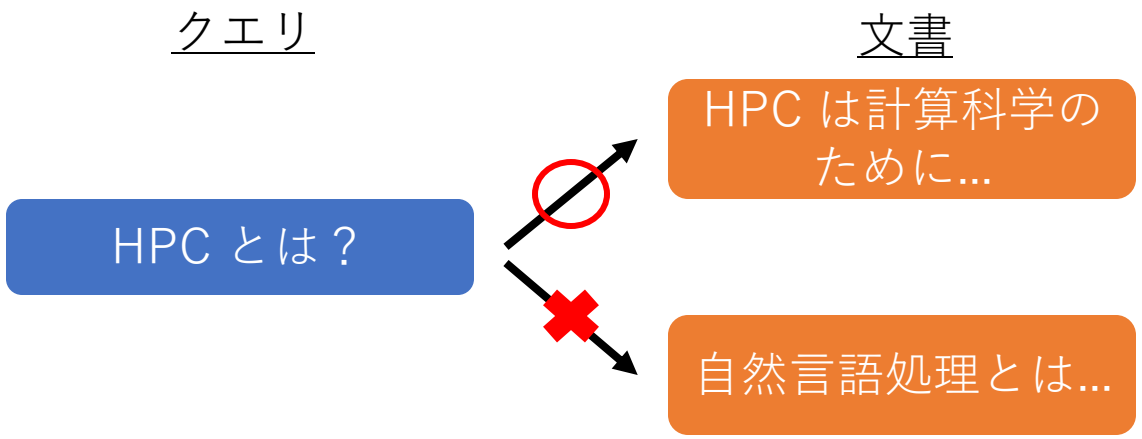


開発用タスク	スコア
タスクA	3.23
タスクB	55.3
タスクC	4.47

モデル構築: 1. 学習データ作成

テキスト埋め込みモデルの学習データとして、「あるクエリに対して、関連した文書、関連しない文書」の事例を集める。
今回の構築では、「既存の自然言語処理タスク」からそのようなデータを作成した。
（詳細は言語処理学会第31回年次大会「インストラクションと複数タスクを利用した日本語向け分散表現モデルの構築」）

学習データの事例概要



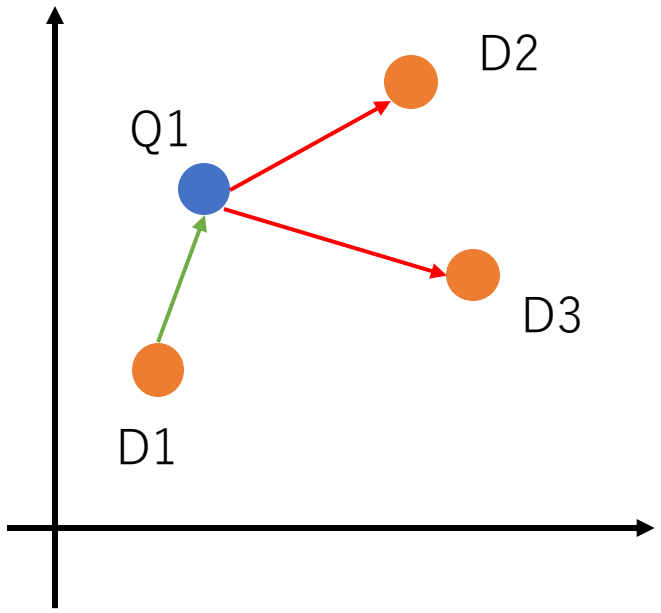
学習データ作成に使用した既存タスクの概要

Question Answering 4 tasks 398K examples	Summarization 2 tasks 4K examples	Natural Language Inference 7 tasks 418K examples	Paraphrase 2 tasks 66K examples
Classification 1 task 19K examples	Machine Translation 4 tasks 551K examples	Retrieval 3 tasks 329K examples	English Tasks 26 tasks 1M examples

モデル構築: 2. 学習方法検討

今回のテキスト埋め込みモデルの学習では、「関連するものを近づけ、関連しないものを遠ざける」学習を実施した。
具体的には、右下図のようにクエリ、関連文書、非関連文書の三つ組から構成されたミニバッチを作成し、これらの関係性を教師信号としてテキスト埋め込みモデルの訓練を実施した。

学習概要図



ミニバッチ例

クエリ	関連文書	非関連文書
HPC とは？	HPC は計算科学のために…	自然言語処理とは…
数理最適化とは？	数理最適化とは…	最適な勉強方法の探し方とは…
…	…	…
動的計画法とは？	DP は対象を部分問題として…	計画的に物事を進める方法は…

モデル構築: 2. 学習方法検討

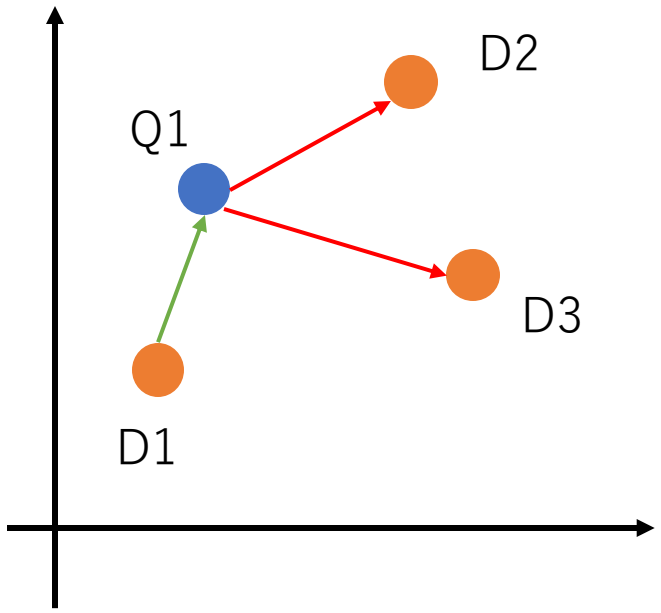
テキスト埋め込みモデルの学習で使われる工夫

- in-batch negative: ミニバッチ内のあるクエリに対して、非関連文書は他のクエリの文書も使用する

→ ミニバッチサイズが大きければ大きいほど、非関連文書（負例）の数を増やせる！

→ VRAM が大きい計算機環境が要求される状況

学習概要図



ミニバッチ例

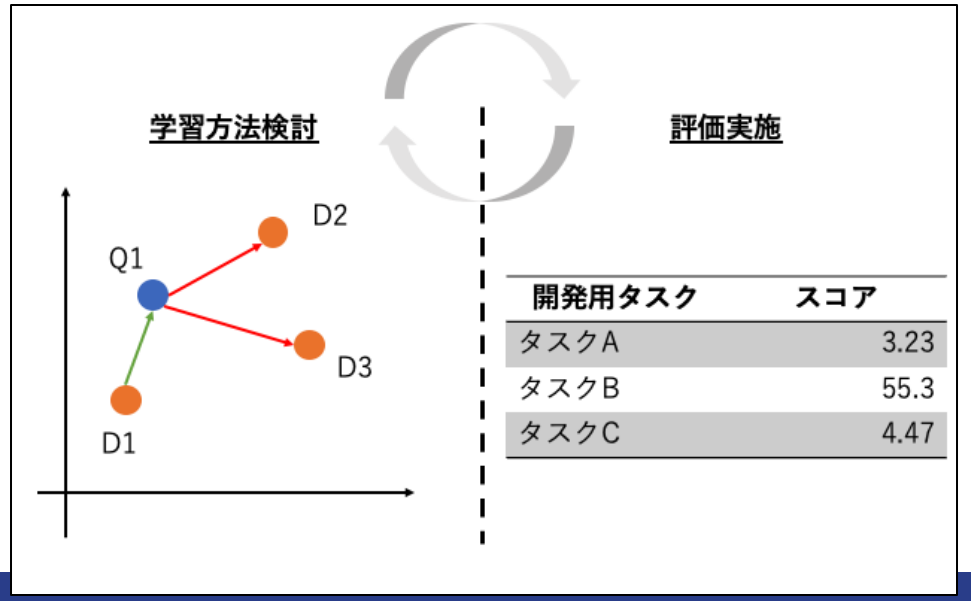
クエリ	関連文書	非関連文書
HPC とは？	HPC は計算科学のために...	自然言語処理とは...
数理最適化とは？	数理最適化とは...	最適な勉強方法の探し方とは...
...	...	...
動的計画法とは？	DP は対象を部分問題として...	計画的に物事を進める方法は...

モデル構築: 2. 学習方法検討

今回探索したハイパーパラメータ

- ベースモデル（事前学習済み BERT）
- 学習データのバランス
- ミニバッチサイズ
- 系列長
- 学習率
- 学習率スケジューラー

→ 各項目について、候補値を決めグリッドサーチで最適なハイパーパラメータを決定
→ 学習・評価の時間はできるだけ短くしたい！



モデル構築: 2. 学習方法検討

玄界でどの程度学習が効率化されたか、RTX A6000 と比較を実施し、その統計値をまとめた。
学習設定は次の通り。

- 学習対象モデルのパラメータ数
  - BERT base サイズ（# 153M; GPU1枚で学習）
- 系列長: 512
- 学習 Epoch 数: 2
- 総合ミニバッチサイズ: 1024
  - 玄界では 1024/GPU
  - A6000 では 512/GPU x Gradient Accumulation 2 step※

	BERT base サイズ	
カテゴリ	玄界環境 b-batch H100 1枚	RTX A6000 1枚
総学習時間 [h]	5.86	14.68
秒間学習ステップ数 [件数]	0.296	0.106

※ A6000 ではミニバッチサイズ1024は Out of Memory となった。

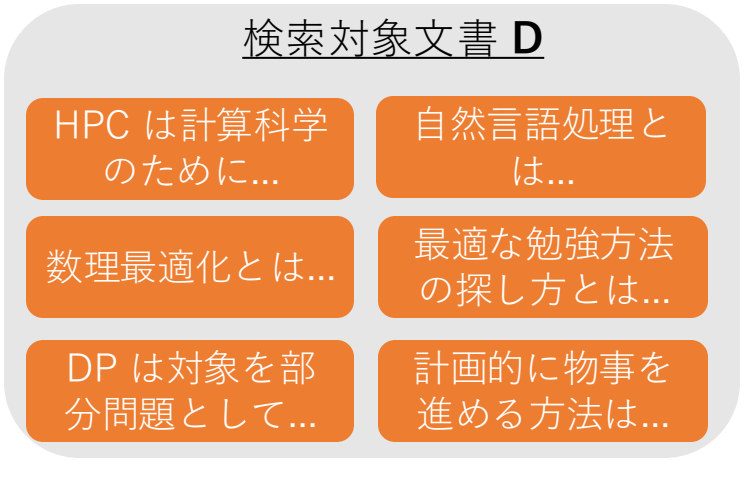
モデル構築: 3. 評価方法実施

開発したモデルの評価は「開発用タスク」「評価用タスク」で実施した。
開発用タスクと評価用タスクは検索タスクを中心に様々なタスクで構築されている。
（例: クエリ数720件、検索対象文書7M件の検索タスク）

検索タスクの実行手順

1. 検索対象文書のベクトル化
2. クエリのベクトル化
3. クエリのベクトルと検索対象文書全てのベクトルの類似度計算
   → 類似度の高い文書を出力

検索対象文書のベクトル化



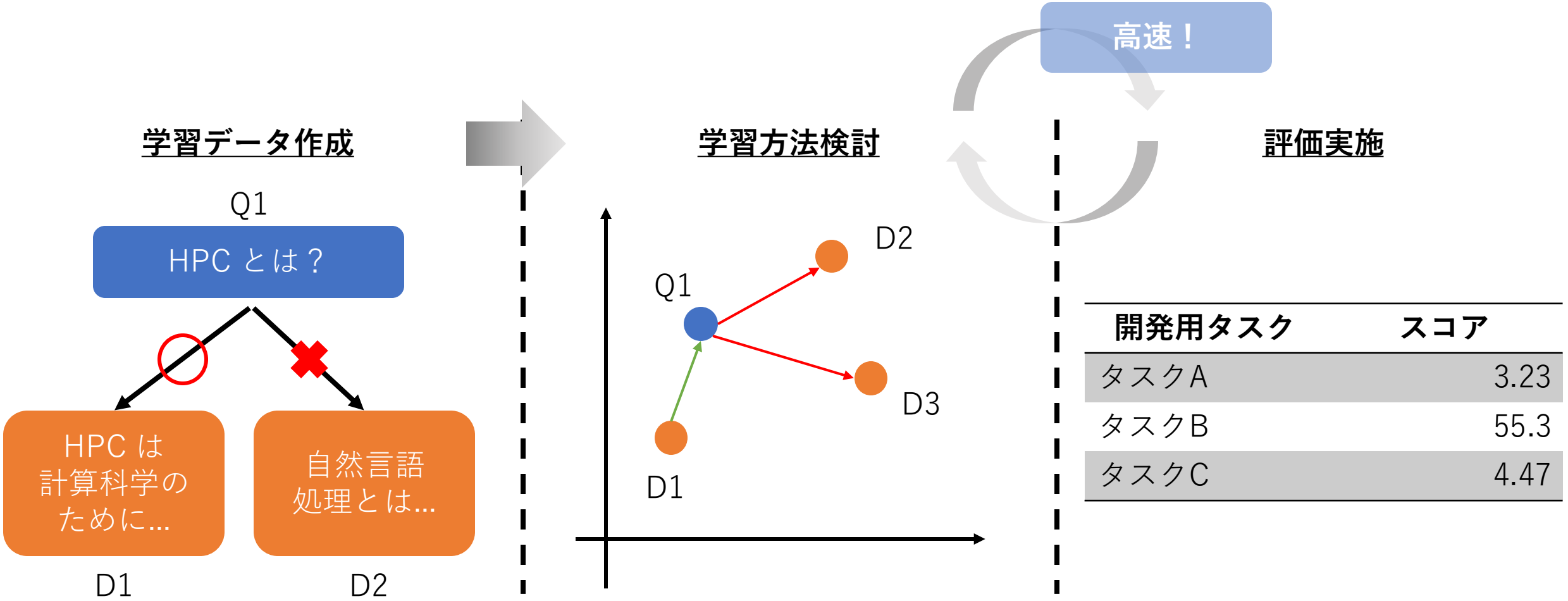
$Vec^{|D|\times dim}$

クエリのベクトル化

Vec^{dim}

テキスト埋め込みモデルの構築に向けた取り組み

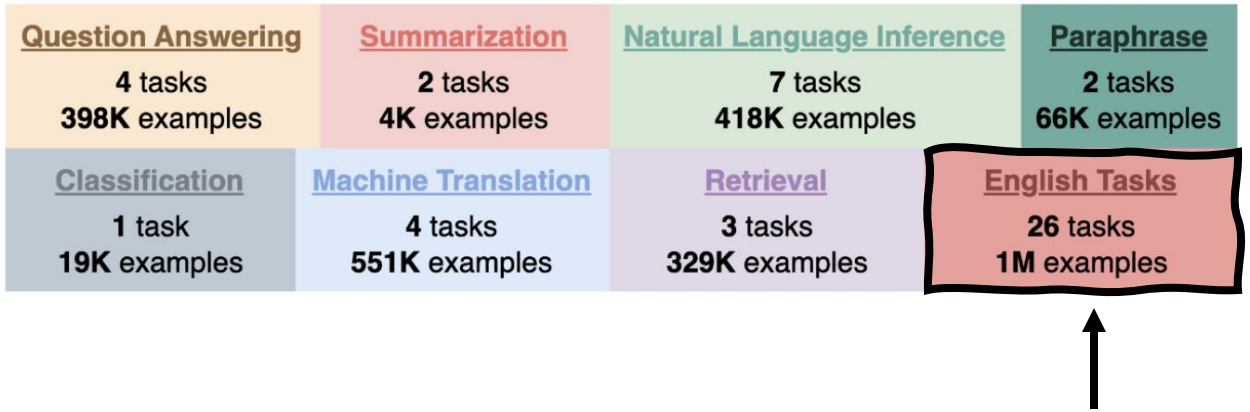
玄界環境を利用することで、学習方法と評価のサイクルを高速に実現することができた。
次ページからは、得られた知見について簡単に共有を行う。

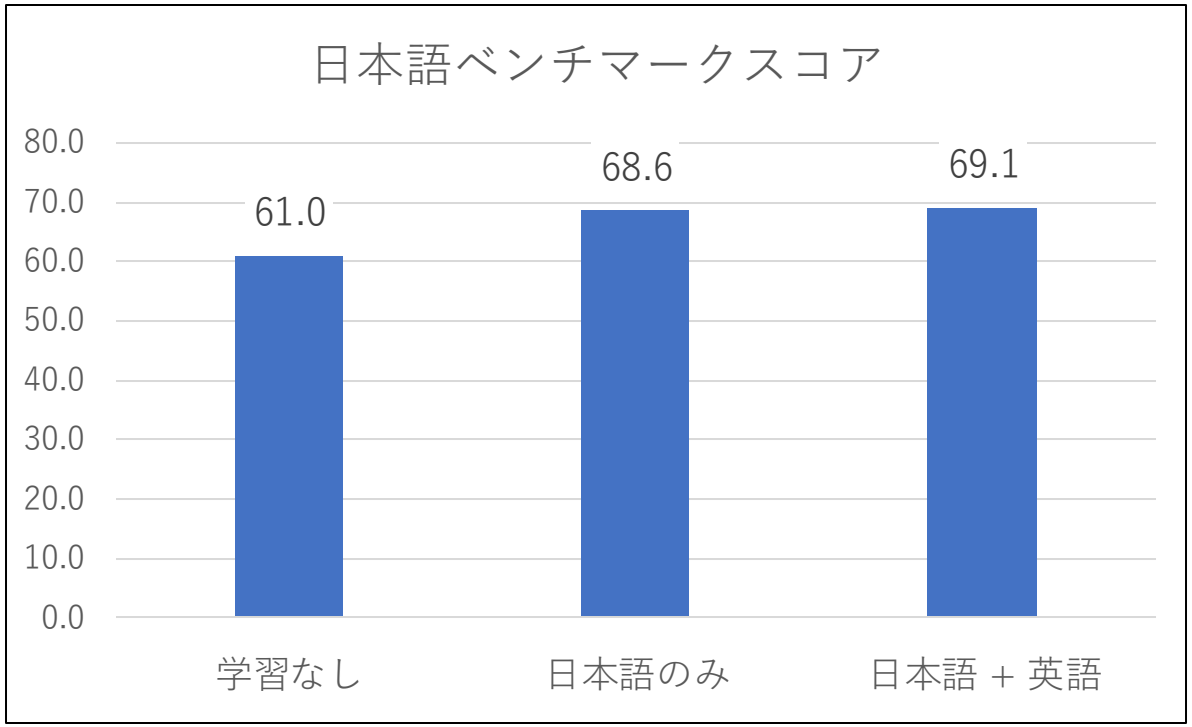


開発用タスク	スコア
タスクA	3.23
タスクB	55.3
タスクC	4.47

実験を通して得られた知見: 学習データの影響

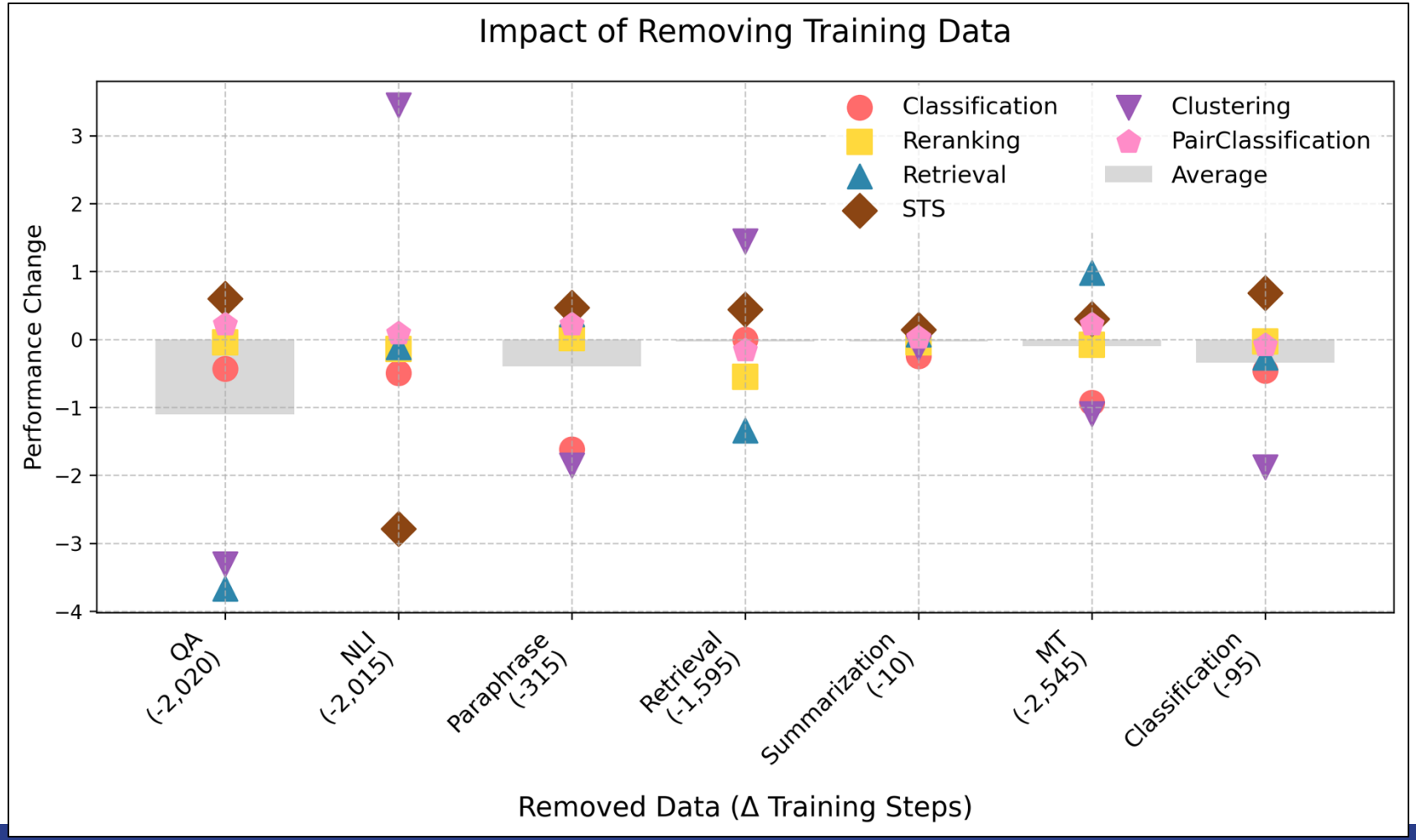
学習データの構成要素として、「英語データを含めることで日本語テキスト埋め込み性能が向上するか」検証を実施した。
→ 英語データを含めることで、日本語性能も向上した！





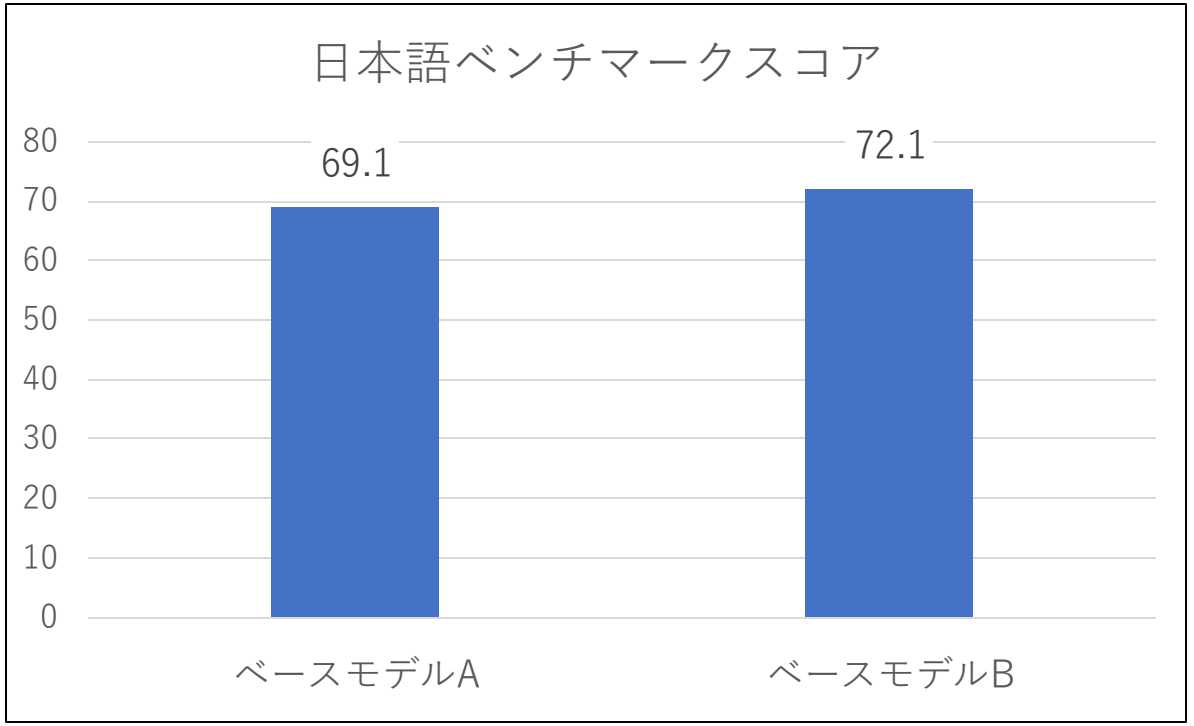
実験を通して得られた知見: 学習データの影響

学習データの構成要素として、「どの種類の学習データかどの評価タスクに効果的か」検証を実施した。
→ テキスト埋め込みを評価するタスクに応じて、効果的な学習データが異なることを確認！



実験を通して得られた知見: ベースモデルの影響

ハイパーパラメータとして、「ベースモデルを変更した際の精度への影響」検証を実施した。
→ ベースモデルに応じて、性能が大きく変わることを確認！



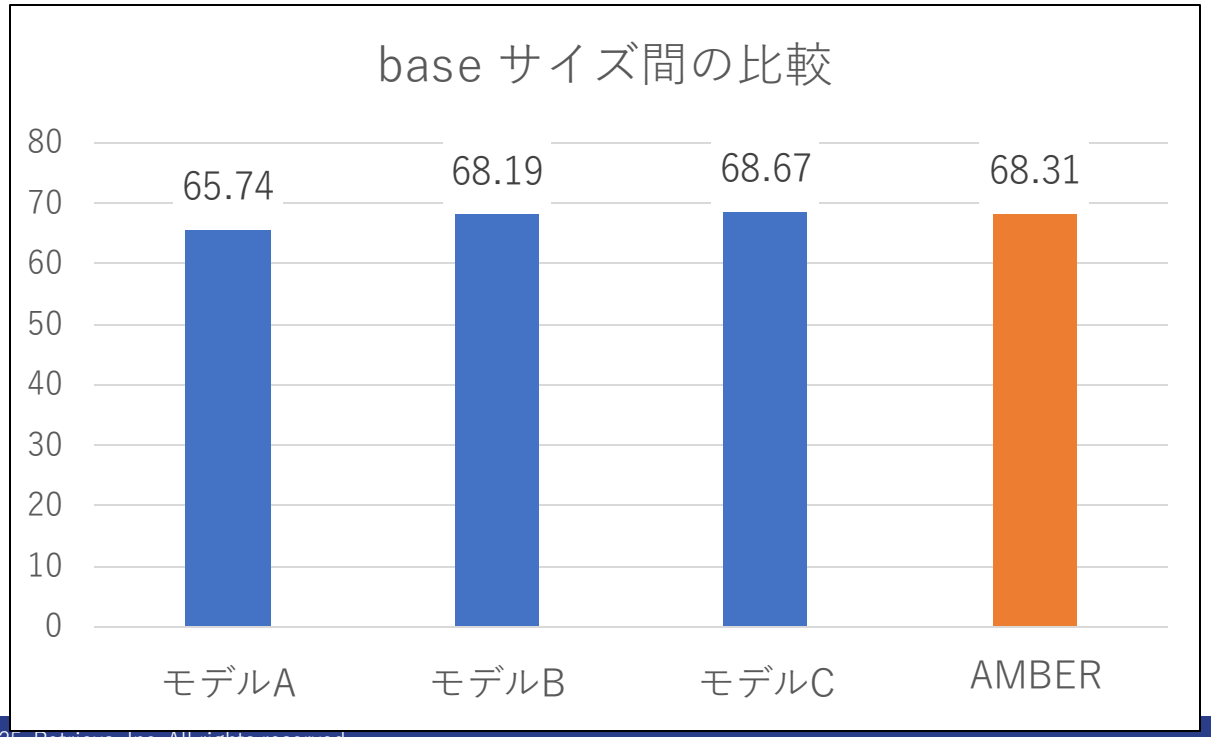
作成したモデルの性能

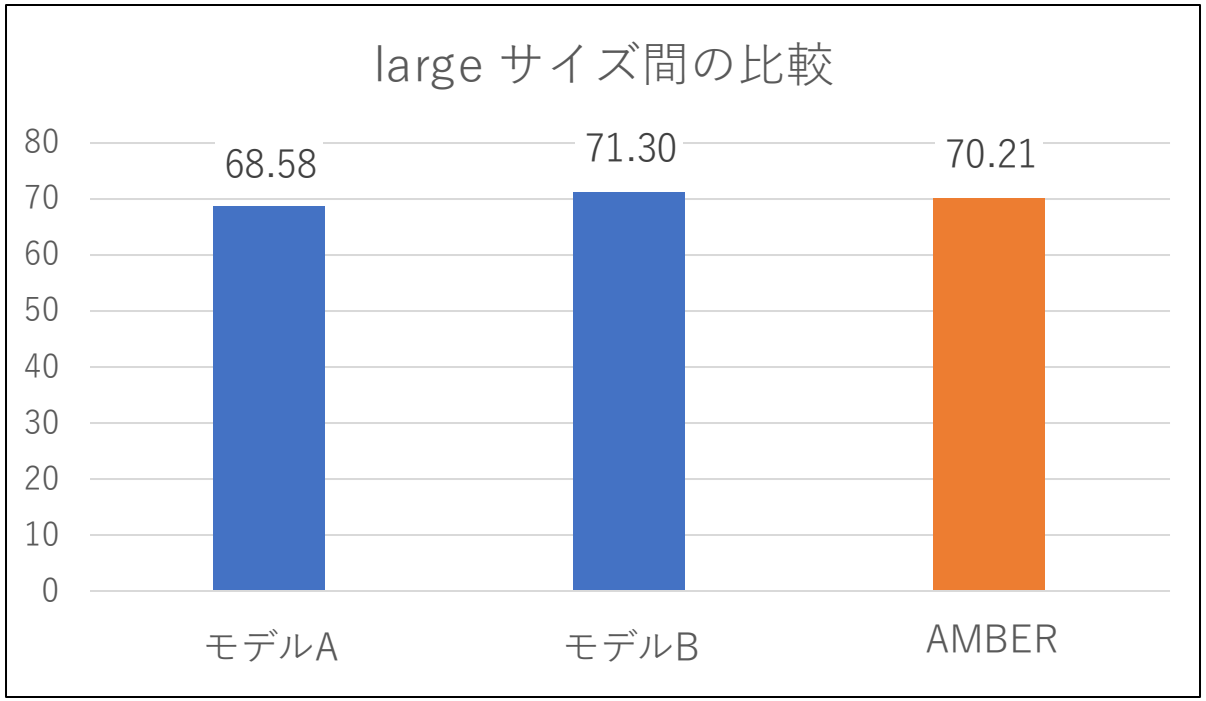
作成したモデルがどの程度の性能か検証を実施するため、日本語検索評価データで評価を実施した。

検証モデル※

- AMBER: 本プロジェクトで作成・公開したモデル（base, large）
- モデルA: cl-nagoya/ruri-base, cl-nagoya/ruri-large
- モデルB: cl-nagoya/ruri-base-v2, cl-nagoya/ruri-large-v2
- モデルC: pkshatech/GLuCoSE-base-ja-v2

※ AMBER 公開時に公開済み、または公開日が近いモデルで比較を実施。





まとめ

- 玄界を利用して「テキスト埋め込みモデル」の構築を実施した
  - **玄界を利用することで、学習速度・評価速度の高速化などの恩恵を受けることができた**
- 複数回の検証を経て、様々な知見を得ることができた
  - **学習データの影響、ベースモデルの影響...**
- 作成したモデルは、既存の埋め込みモデルと遜色ない性能であることを確認した
  - **本モデルは huggingface-hub にて公開中！**
    - retrieva-jp/amber-base
    - retrieva-jp/amber-large

会社名	株式会社レトリバ
設立	2016年8月（9期目）
所在地	東京都豊島区西池袋1-11-1 メトロポリタンプラザビル14F WeWork内
従業員数	44名
代表	田口琢也
事業内容	自然言語処理及び機械学習を用いた ソフトウェアの研究・開発・販売・導入およびサポート
グループ企業	株式会社万葉

レトリバ