LLM-jp モデルを用いた Fine-tuning によるドメイン特化 Embedding モデルの構築

高木純平[1)], 村中勇輝[1)], 小林伸次[1)], 浦上暉允[1)], 河野聖[1)]
1) チームS (株式会社セック)

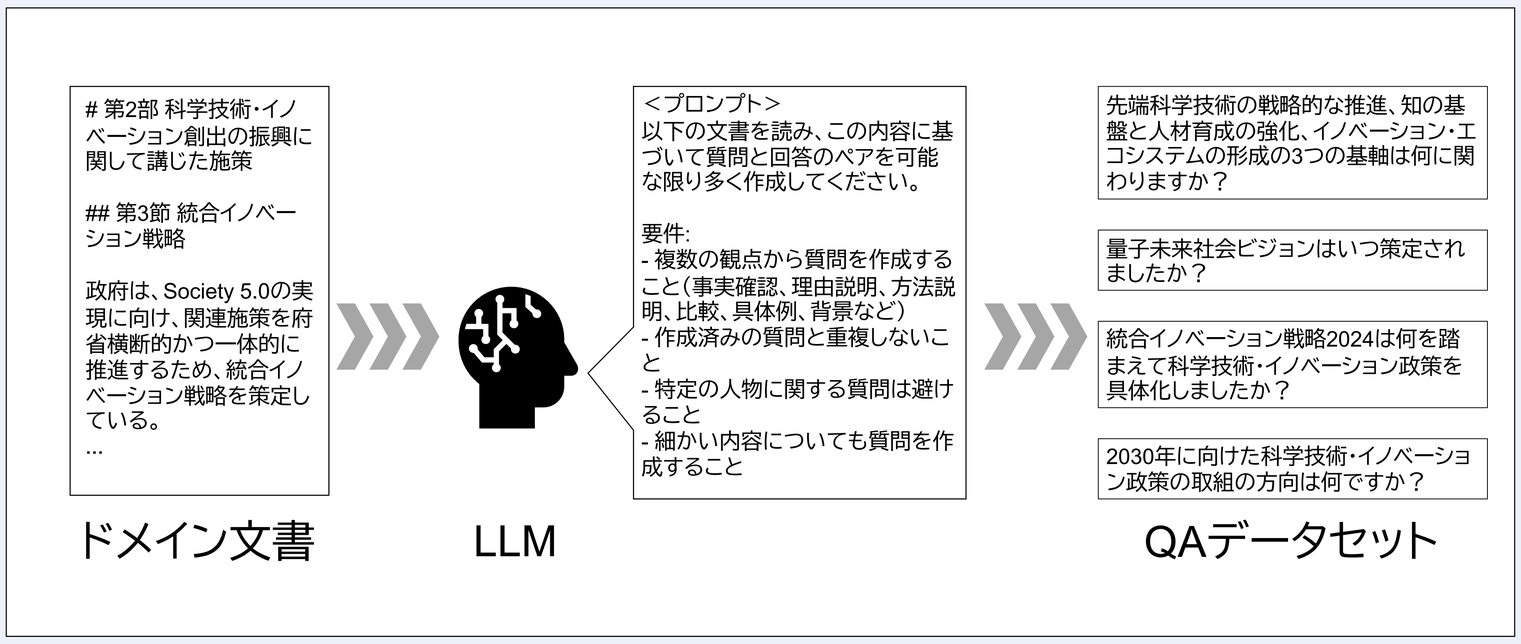
■ 背景・アプローチ

RAG の性能は前段の検索品質に大きく依存するが、汎用 Embedding モデルでは専門ドメインにおいて意味検索精度が不足し、回答品質・速度・コストの面でボトルネックとなるケースが多い。対象ドメインに特化した RAG の回答精度・速度・コストを最適化するために、専用 Embedding モデルを作成する。Fine-tuning 用データセットの生成には、日本語性能が高く指示追従性に優れる LLM-jp を用いることで高品質な QA データを安定的に得られると考え、これを使用した。生成した特定ドメイン向け日本語 QA データセットにより、Embedding モデルの効果的な Fine-tuning を行う。

■ データセット作成

想定質問を LLM で作成し、
(質問, 正解チャンク)のデータセットを作成。

- 対象ドメイン
  - 令和7年版_科学技術・イノベーション白書
- 生成モデル
  - llm-jp / llm-jp-4-8b [v4-8b-decay2m-ipt_v3.1-instruct4]
- 作成されたデータセット
  - チャンク数: 204
  - 学習データ: 21,659 個の想定質問



■Fine-tuning による検索性能評価

Overview

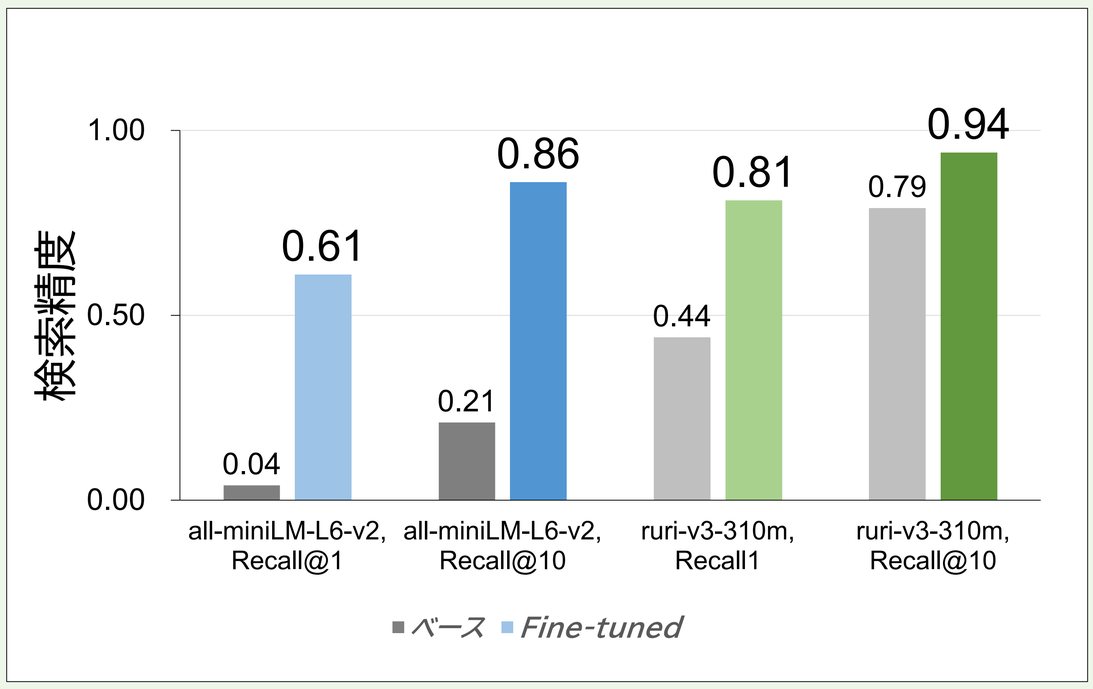
- Fine-tuning により、ドメイン知識の検索性能が大幅に上昇した。
- all-miniLM-L6-v2 のような軽量モデルでも、Fine-tuning 後ではドメイン知識検索においてベースモデルの ruri-v3-310m を上回る検索性能を示した。

Fine-tuning

- ベースモデル
  - sentence-transformers / all-MiniLM-L6-v2（22.7M）
  - cl-nagoya / ruri-v3-310m（315M）
- 損失関数
  - **Multiple Negatives Ranking Loss**
- 代表的な学習パラメタ
  - batch: 128、epoch: 100

Result

- 評価には、openai / gpt-oss-20b を使用し、上記の「データセット作成」と同様の方法で作成した 6,115 問のデータセットを使用した。
- Recall@1、Recall@10で Fine-tuning 前後の検索性能の比較を行った。
- all-miniLM-L6-v2では、Recall@1、Recall@10 ともに**60% 以上の大幅な上昇**を示した。検索性能が既に高い ruri-v3-310m においても、**Recall@1 では 37% 、Recall@10 では 15%** の上昇を示した。



■ Rerankとの比較

Overview

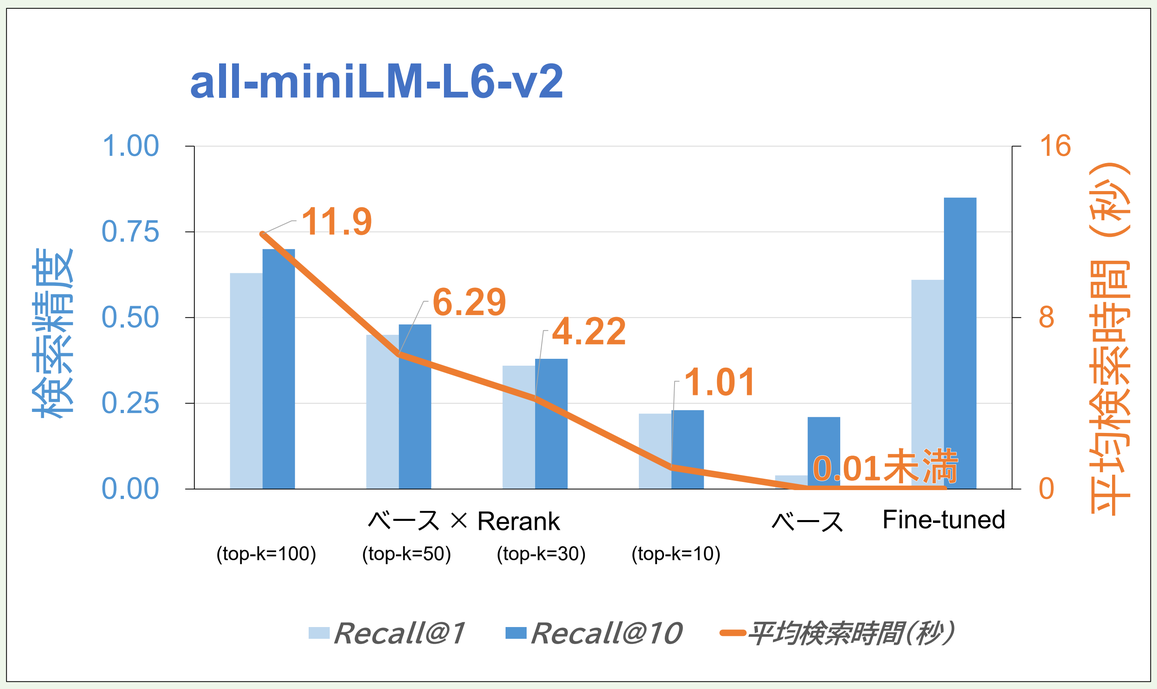
- Fine-tuning を行ったモデルは、ベースモデルと同等のミリ秒オーダーの検索速度を維持しながら、Rerank で再順位付けした場合と同程度の検索性能を示した。

Rerank

- Rerank モデル
  - BAAI / bge-reranker-v2-m3（0.6B）

Result

- Recall@1、Recall@10でベースモデル × Rerank の精度と、Fine-tuning のみモデル検索精度の比較を行った。
- Fine-tuning を行ったモデルは、ベースモデルによるランキングの上位 30 件、50 件を用いて再順位付けした結果と**同程度またはそれ以上の検索性能**を示した。
- Fine-tuningを行ったモデルは、精度の高い検索を**ミリ秒オーダーの検索時間**で行うことが可能である。



■ まとめ

LLM-jp で生成したデータセットを活用することで、特定ドメインに適した効果的なモデル学習が可能であることを示した。特筆すべきは、Fine-tuning 済みの軽量モデルが、Fine-tuning を行っていない高性能 Embedding モデルを凌駕する検索性能を発揮した点である。これは、限られた計算リソースでも高速かつ高精度な検索システムが実現可能であることを示すとともに、リランキング等の後処理手法と組み合わせることで、さらなる性能向上も期待できると考えられる。