

理研BRCにおける知識グラフの記述文 (Description) データに対する検索手法の評価

櫛田達矢、臼田大輝、高田豊行、梶屋啓志（理研BRC）

理化学研究所バイオリソース研究センター（理研BRC）では、保有する実験動物・植物、細胞、遺伝子、微生物株の情報を、RDFベースの知識グラフ（KG）として一元管理している。現在、このKGに対しては、SPARQL検索のほか、リソース番号や遺伝子名などを用いた「バイオリソース検索」(<https://web.brc.riken.jp/>)、疾患や表現型オントロジーを利用した「アドバンスド検索」(<https://knowledge.brc.riken.jp/advanced>)が提供されている。しかし、これらの既存手法では、技術的な制約から各リソースの説明文（Description）の情報が十分に活用されていない。Descriptionは、関連する変異遺伝子、特異的発現組織、表現型などの情報が日本語と英語で簡潔に記述されている文章データであるが、単語レベルでの全文検索ができないため、その潜在的な価値を十分に引き出せていない。本発表では、このDescriptionデータに対する有効な検索手段を確立することを目指し、マウスのデータを用いて実施した全文検索エンジン、ベクトルデータベース、大規模言語モデルのパフォーマンス評価の結果を報告する。

キーワード：知識グラフ（KG）、全文検索エンジン、ベクトルデータベース、大規模言語モデル（LLM）、Retrieval Augmented Generation (RAG)

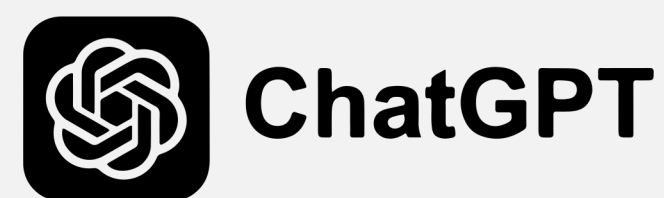
AIプラットフォームを使った新規検索システムの可能性の調査概要

目的

- AIプラットフォームを活用してバイオリソースの取得システムを強化し、Q&Aインターフェースとファジー検索機能を実装する。

技術要素

- 全文検索エンジン: elasticsearch
- ベクトルデータベース: Faiss (Facebook AI Similarity Search)
- LLM: gpt-3.5-turbo (ChatGPT API)
- Retrieval Augmented Generation (RAG)
- GoogleTranslator



使用するデータ

- マウス系統の説明文



機能

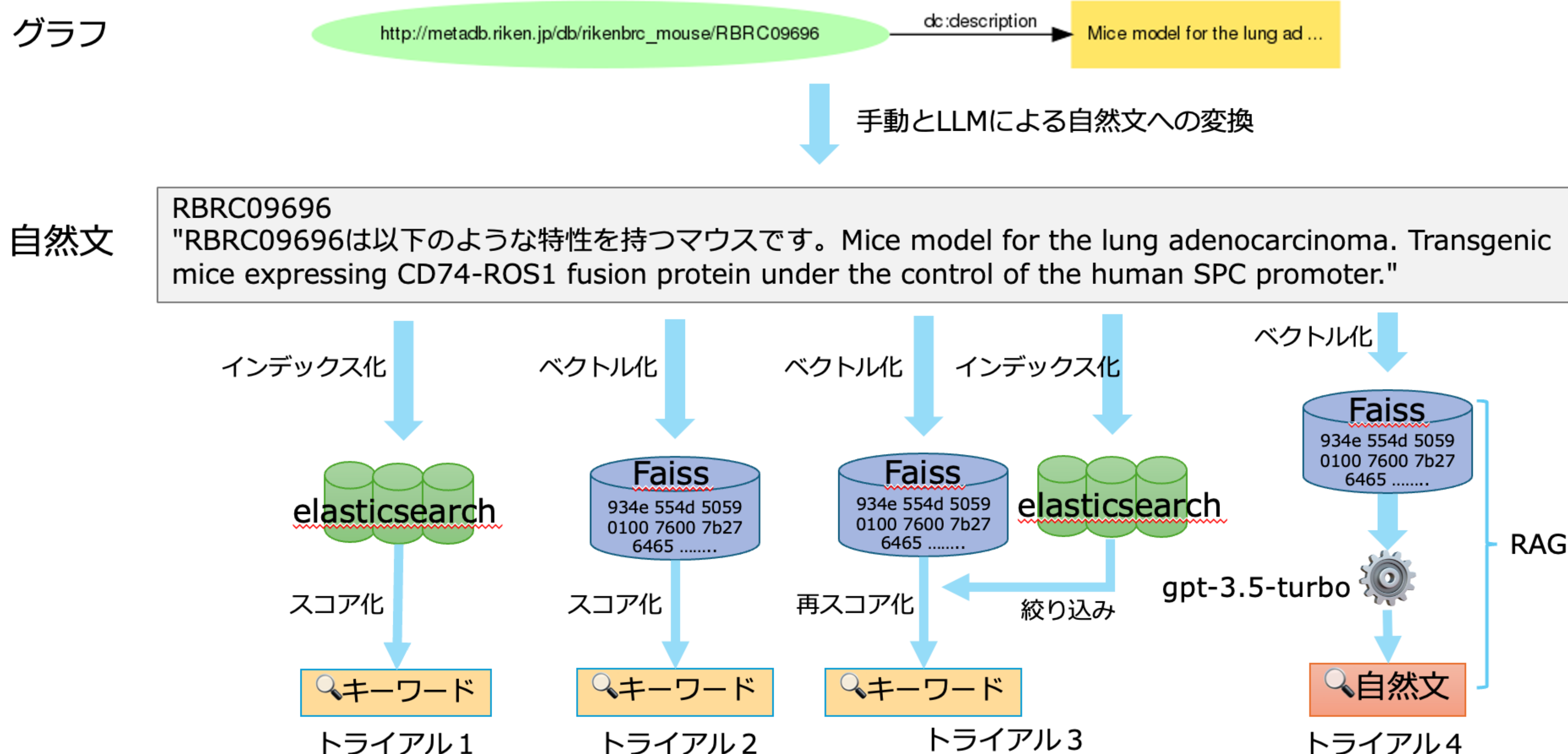
- 日英翻訳機能付きの対話検索インターフェース。
- 特定の用途（例えば病気の研究など）に適したマウスを探索し、推薦する。



期待する効果

- ユーザーがより直感的で柔軟なクエリを実行できるようにする。
- 全体的なUX（ユーザーエクスペリエンス）を向上させる。
- ユーザーによる入力時のミススペルによる検索の失敗を防ぐ。
- 検索精度（特に、再現率）を向上させる。

AIプラットフォームを用いたグラフデータ処理の例



トライアル1-4の評価

開発者による評価

トライアル No.	検索タイプ	入力する語句の例	全文検索エンジン	ベクターDB	LLM	評価 (ヒト)
トライアル1: Elasticsearch	キーワード	lung cancer, RBRC02435, tumor model など	elasticsearch	-	-	★★★☆☆
トライアル2: ベクトルDB			-	Faiss	-	★★★☆☆
トライアル3: Elastic search + ベクトルDB			elasticsearch	Faiss	-	★★★★☆
トライアル4: ベクトルDB+ ChatGPT (RAG)	自然文	Which mice are associated with human lung cancer?	-	Faiss	gpt-3.5-turbo	★★★★☆

ChatGPT 4oによる評価

トライアル No.	Proposed System Name	Feature	Evaluation (LLM)	Notes on Evaluation
トライアル1: Elasticsearch	TextMouseQA	Full-text search using Elasticsearch	★★★☆☆	Basic keyword matching, lacks flexibility in handling paraphrased or abstract queries.
トライアル2: ベクトルDB	VecMouseQA	Semantic search using vector embeddings (e.g. E5)	★★★★☆	Captures semantic similarity, but lacks exact matching control.
トライアル3: Elastic search + ベクトルDB	HybridMouseQA	Combined search using Elasticsearch and FAISS	★★★★☆	Balances precision and semantic recall effectively.
トライアル4: ベクトルDB+ ChatGPT (RAG)	RAG-based MouseQA	LLM-based QA (ChatGPT) with vector + keyword context	★★★★★	Offers natural language answers and better reasoning, though dependent on model behavior.

トライアル4（QA with LLM）の質疑応答例

🧠 質問を入力してください（終了したいときは 'exit' と入力）
> Which mice are associated with human lung cancer?

💬 回答:

Mice with RBRC09696 and RBRC09697 are associated with human lung cancer. This is because these mice are model mice that develop lung cancer and are transgenic mice expressing fusion proteins discovered in human lung adenocarcinoma. Specifically, RBRC09696 expresses CD74-ROS1 and RBRC09697 expresses SDC4/ROS1, and these fusion proteins have been reported to be associated with human lung adenocarcinoma. Therefore, these mice can be considered important models in the study of human lung cancer.

RBRC09696およびRBRC09697を発現するマウスは、ヒト肺がんとの関連性が指摘されています。これは、これらのマウスが肺がんを発症するモデルマウスであり、ヒト肺腺がんにおいて発見された融合タンパク質を発現するトランスジェニックマウスであるためです。具体的には、RBRC09696はCD74-ROS1を、RBRC09697はSDC4/ROS1を発現しており、これらの融合タンパク質はヒト肺腺がんとの関連性が報告されています。したがって、これらのマウスはヒト肺がん研究における重要なモデルと考えられます。

