統合データベース検索におけるテキストエンベディング利活用

千葉啓和、藤原豊史、守屋勇樹、池田秀也、申在紋（ライフサイエンス統合データベースセンター）

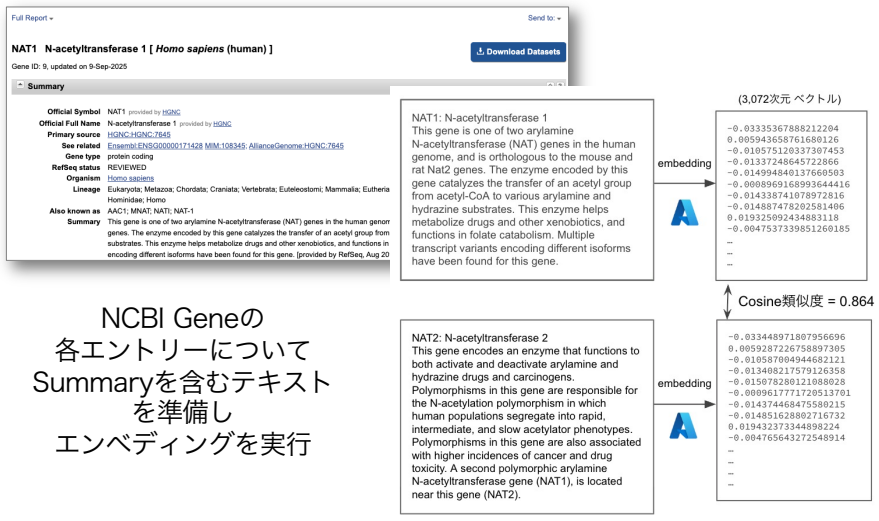
Motivation

バイオメディカル分野のデータベースの中には、構造化された情報だけでなく、自由記述のような非構造化テキストも含まれている。
それらのテキストに対し、テキストエンベディングモデルを適用することによって、意味的な類似性に基づく検索を可能にする。
異なるカテゴリのデータを同じベクトル空間にマップすることにより、カテゴリをまたいだ統合的な検索も可能になると考えられる。

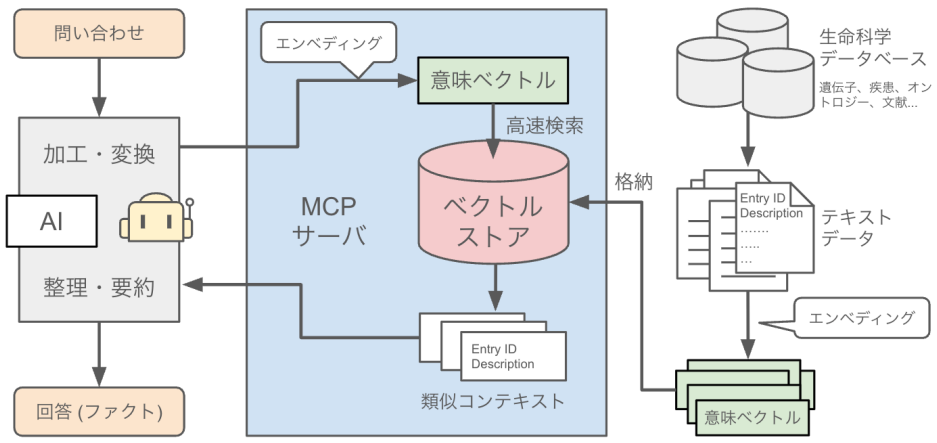
Methods

非構造化テキストの意味的類似性を表現するため、OpenAIの text-embedding-3-large を用いた。また、ベクトル検索の結果を整理・要約するために、OpenAIの gpt-4o を用いた。これらのモデルは、Microsoft Azure上にデプロイした。ベクトルストアとしてFAISSを用いた。

テキストエンベディングによって得られるベクトルの例



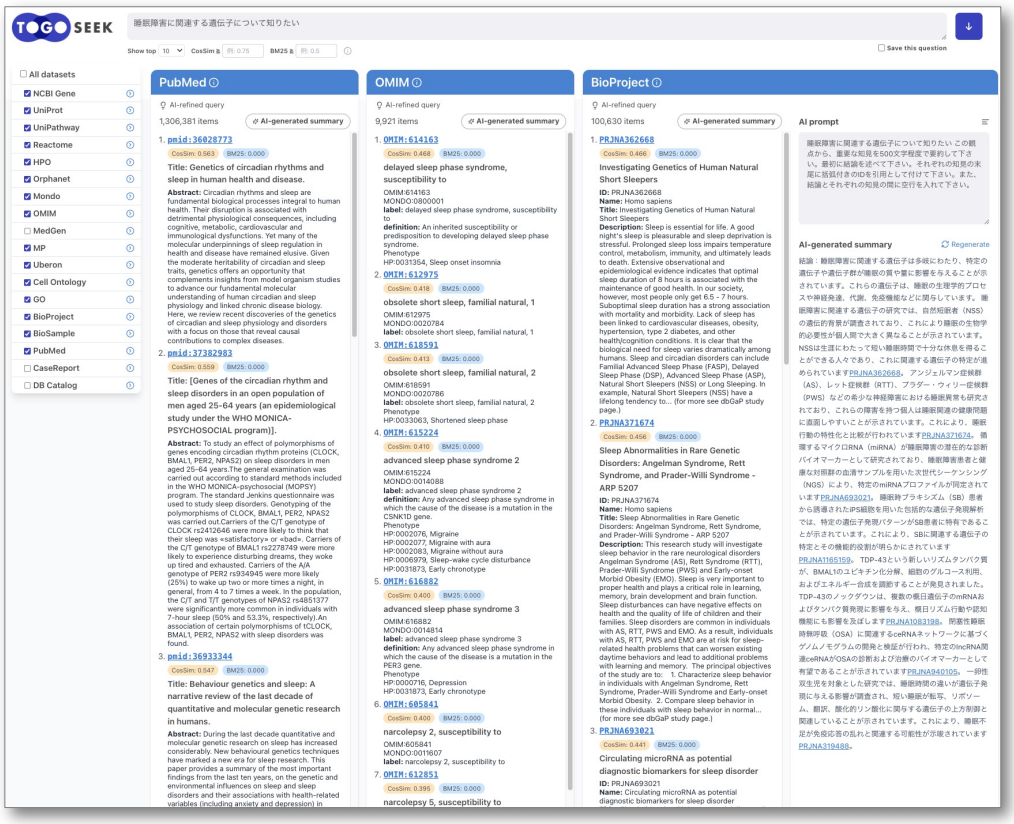
テキストエンベディングを用いた統合データベース検索システムの概要



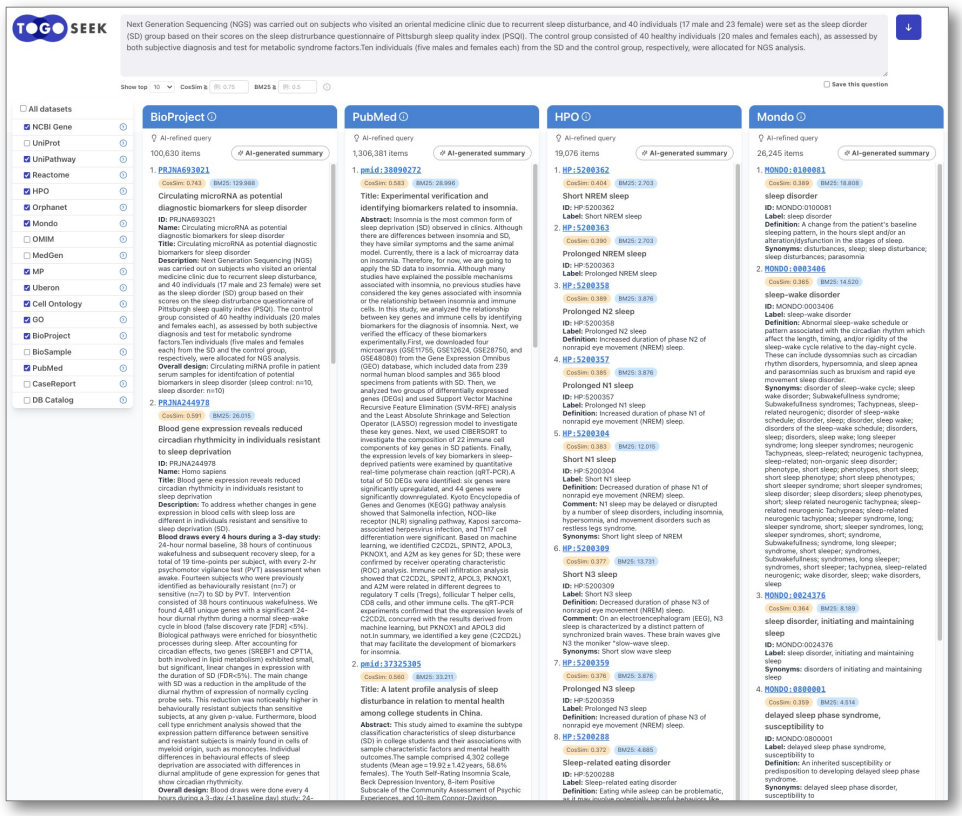
ライフサイエンス統合データベースセンター（DBCLS）では、大規模言語モデル及びエンベディング技術を用いて、データベースと非構造化データを統合的に利用するための技術開発を進めている。特にここでは、テキストエンベディングを活用してデータベースを検索するシステムの開発について発表する。本システムでは、バイオメディカル分野のデータベースに含まれるエントリーを対象として、事前にテキストエンベディングを実行し、各エントリーの埋め込み表現を生成する仕組みを構築している。さらに、整備済みのデータセットに対して、ユーザーが入力した自然文の埋め込み表現に基づくベクトル検索を行い、検索結果を要約して提示することができる。検索結果を要約する際に、データベースエントリーをファクトとして付加することにより、ユーザーが入力した文章に関するファクトチェックを支援することもできると考えている。本発表では、開発中のシステムについて紹介したい。

Example Use Cases

ユーザーが入力した自然文と意味的に類似性の高いエントリーを、さまざまなデータセットに対して横断的に検索することができる。
使用したモデルは多言語に対応しているため、入力は日本語でも英語でも良く、また、単語の列だけでなく文や段落でも良い。



入力されたクエリをエンベディングし、ベクトル検索によって意味的類似性の高いエントリーを抽出。さらに、それらの結果をLLMにより整理・要約する。



研究概要を記したパラグラフを入力して、関連する研究を探すことができる。関連するオントロジーのタームも得られる。アノテーション支援にも応用できると考えられる。