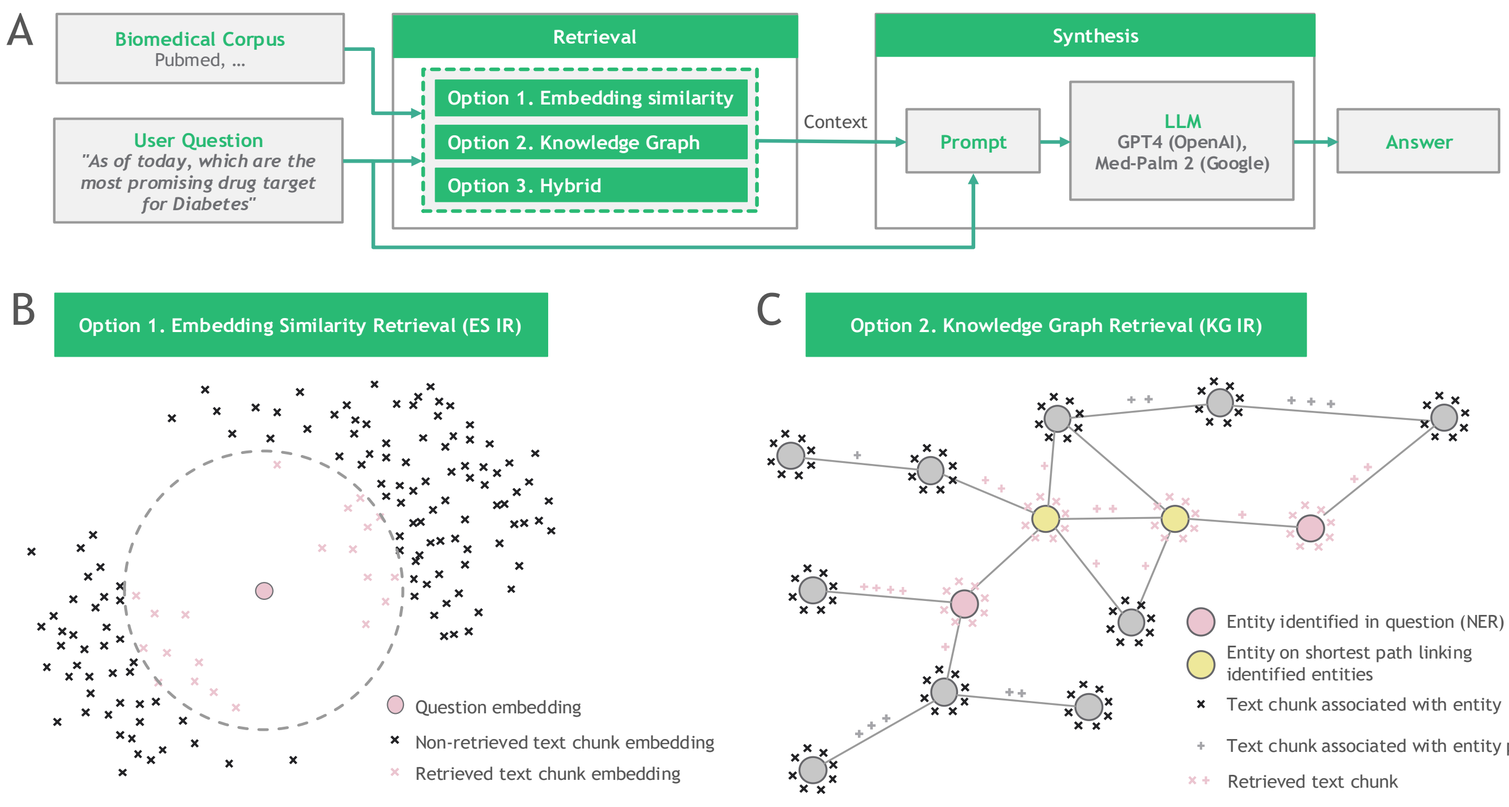
Graph-Based Retriever Captures the Long Tail of Biomedical Knowledge

Julien Delile, Srayanta Mukherjee, Anton Van Pamel, Leonid Zhukov

The Boston Consulting Group

In this study we introduce a novel knowledge-graph-based retrieval approach that enables access to the long tail of biomedical knowledge. We demonstrate that simple RAG retrieval approaches leave out a significant proportion of relevant information when over-represented topics monopolize the list of most similar text chunks. We propose to perform a rebalancing of the retrieved text chunks by under-sampling larger clusters of information through structuring the text corpus with a Knowledge Graph (KG) of biomedical entities (genes, diseases and diseases). In addition, our method also provides control mechanisms to prioritize the retrieval of recent and impactful discoveries. Finally, we built a hybrid approach combining the strengths of LLM embedding semantic relationships and structured knowledge graph and show that it outperforms both embedding similarity (ES) and KG based methods for biomedical information retrieval (IR).

Retrieval pipeline used in experiments



Retrieval

Embedding Similarity retrieval (ES IR)

- Use cosine similarity to rank the embedded text chunks for each query.

Knowledge graph retrieval (KG IR)

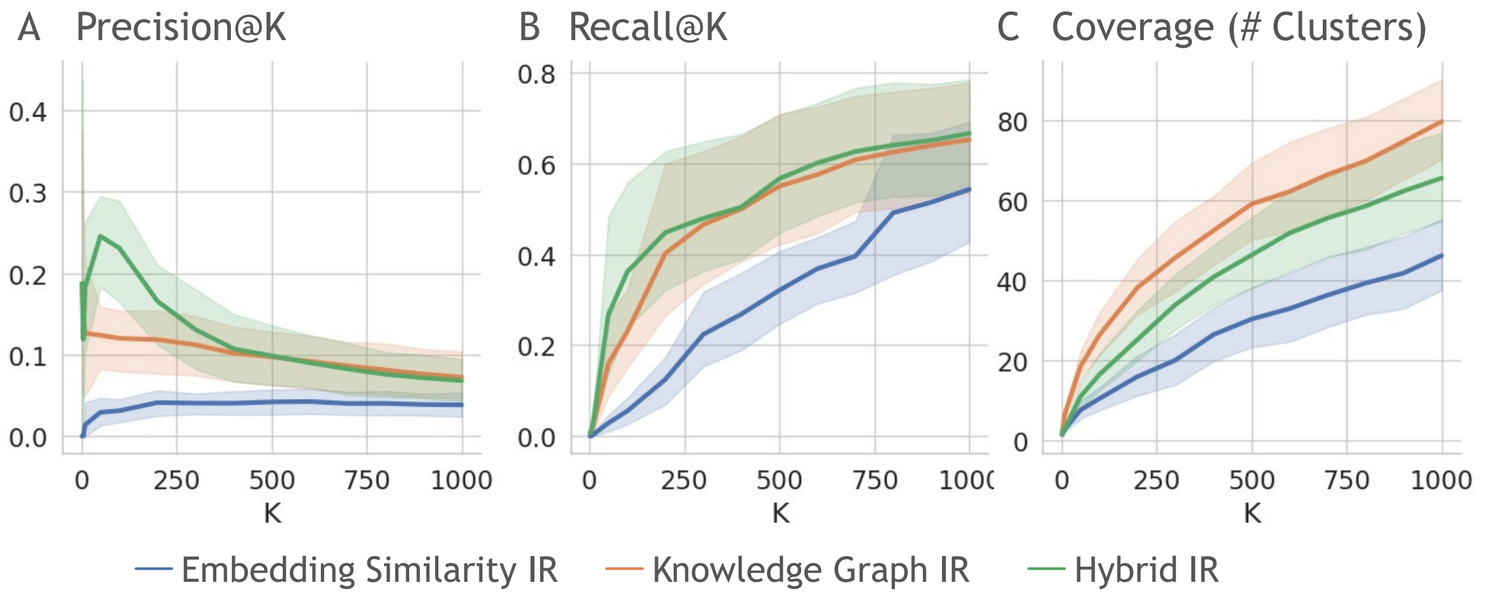
- Identify which entities present in the user question using KAZU pipeline.
- Build the shortest path relating these entities in the graph and retrieve text chunks mapped to the shortest path entities and their neighboring edges.
- Prioritize the most relevant text chunks by a scoring metric that factors in both the recency and the impact of a text chunk using Pareto front of the recency-impact space. The impact of a text chunk is measured as the total number of citations of the associated document.

Hybrid retrieval

- Use ranking score averaging the min-max-normalized ES IR and KG IR scores.

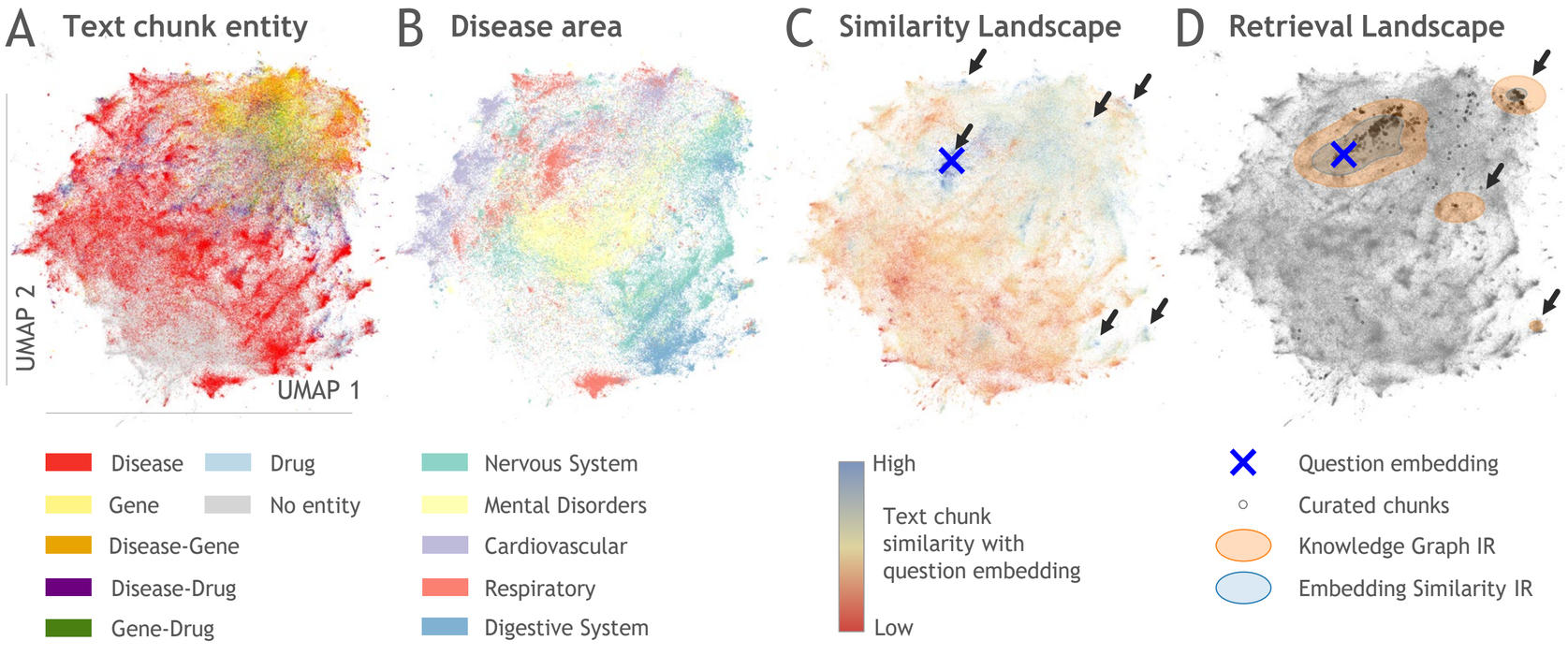
Experiments

- We produced both embedding and KG indexes using Pubmed articles from five therapeutic areas: nervous system, respiratory tract diseases, digestive system diseases, cardiovascular diseases, mental disorders (1% of the articles were sampled randomly)
- We used an open question that requires to exploration a wide range of documents: "What are the known drug targets for treating <disease>?"
- We compared the retrieved information of both approaches with curated annotations produced by biomedical experts containing both an annotation for the disease and for at least one gene in its full text
- We repeated the experiment over 8 diseases selected to cover the different therapeutic areas



Solid lines indicate the metric averages and transparent ribbon 95% confidence interval. KG IR outperforms ES IR on both metrics, with Hybrid IR performing the best.

Discussion



A. UMAP projection of 731k 1536-dimensional embedding vectors
B. Color-coded regions linked to the five covered disease areas and the various types of entities stored in the text chunks that shows localized disease areas in embedding space
C. Embedding of one of the questions used in experiments: "What are the known drug targets for treating Asthma?" and text chunks color-coded by the cosine similarity to the question
D. Top-200 text chunks for both ES IR and KG IR methods, Gaussian kernels over the UMAP coordinates of the retrieved text chunks and the higher density regions visualized by filled contours

The key observations

- Comparing the region of ES IR retrieved text chunks (blue region in Fig. 2D) and from the distribution of gold-standard embedding (grey dots), we observe that the ES IR retrieval region is densely localized in the vicinity of the question embedding.
- KG IR retrieval region is multipolar, covers a wider range of curated documents and allows it to go beyond the immediate surrounding of the question neighborhood to retrieve relevant information capturing the long-tail knowledge of biomedical information.
- ES and KG IR are highly complementary