SRASearch - an enriched NCBI SRA search with a transformer-based neural network

Aadit Kapoor, Aarohi Chopra, Wendy Lee

Department of Computer Science, San Jose State University, San Jose, CA

Keywords: Natural Language Processing, Neural Networks/Transformers, High-throughput Sequencing

ABSTRACT

Biomedical text understanding and retrieval is an active field of research. Advances in genome sequencing technologies have led to an exponential increase in genome sequence data deposited in the public repository, such as the NCBI Sequence Read Archive (SRA). Despite having a vast amount of publicly available data for researchers to share and analyze, the existing NCBI SRA search capability is limited since it lacks the inclusion of semantics in its search functionality. Recent advances in Natural Language Processing (NLP) and Information Retrieval (IR) have allowed us to build search systems capable of understanding search terms' semantics. The influx of publicly available data and NLP advances, such as Bidirectional Encoder Representations from Transformers (BERT) has allowed us to fine-tune our model on applicable downstream tasks such as text classification and Named Entity Recognition (NER). Here we present a search tool, SRASearch, that utilizes NLP techniques to improve search functionality from the NCBI SRA database.

BACKGROUND

Sequence Read Archive (SRA) hosts one of the largest repositories for high throughput sequencing data. These datasets contain multiple tables (known as metadata) that comprise of important information about an experiment.

This metadata is in the form an Extensible Markup Language (XML) submission that maps to a unique accession for each object. These include:

- **Study Name and Abstract**
- **Submission**
- **Sample**
- **Library: Metadata about the sample**
- **Run**

Previous work

- Zhu et al . (2013) developed a R package (SRAdb) capable of querying SRA metadata using SQL (Structured Query Language).
- Choudhary et al. (2019) developed a Python package (pysradb) to query SRA meta using a Python-SQL Connector. An interesting thing to note is that pysradb is an alternate to SRAdb.

TERMINOLOGIES

- **FAISS:** It is a library created by Facebook AI used for efficient similarity search.
- **BLURB:** Biomedical Language Understanding and Reasoning Benchmark is a broad coverage benchmark for PubMed-based biomedical NLP applications containing several datasets and diverse biomedical tasks.

MOTIVATION

What is the Need?

Existing NCBI SRA Search does not take into account the following:

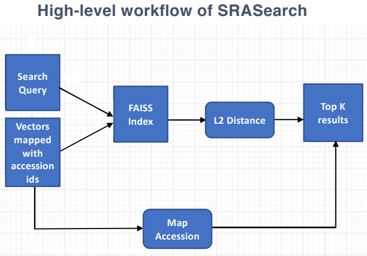
- **Context**: the contextual understanding of the query text such as when we search for "PCR", we get results that contain "PCR Free"
- **Lack of utilization of the SRA metadata**
- **Inaccurate results affect research**

SRASearch aims to solve this problem by:

- **Contextual Understanding**: Incorporate a contextual understanding of the query and reference texts by utilizing various Biomedical NLP techniques.
- **Deeper Search**: SRA (Sequence Read Archive) consists of sub-tables (metadata) that present precious information that can speed and refine the search. Using techniques in NLP (Natural Language Processing), we convert these texts into computable vectors, making them searchable.
- **Open Source**: Utilization of open source software makes the platform accessible and customizable.

METHODS

High-level workflow of SRASearch



We explore the following methods in detail:

- **Keyword Search:** For each query, we employ a brute force method to match each query with the reference query (metadata).
- **One Hot Encoding:** We form one hot encoding for the query and the references (metadata) and then employ cosine similarity.
- **Neural Network Based:** We convert each query (metadata) into a computable biomedical vector. We then perform cosine similarity with the query embedding and the reference embeddings.

A summarization of all the methods highlighting advantages and disadvantages

Method	Speed	Pros	Cons
Keyword Search	Slow	Easy to implement, Requires no preprocessing	Slow Inaccurate results
One hot encoding	Medium	Easy to Implement, Requires Preprocessing of texts	High dimensionality
Neural Network based	Fast	Fast, Heavy preprocessing required	Memory Intensive Index updation required

WORKFLOW

STEP 1 - Preprocessing: Each submission accession (and its corresponding metadata such as abstract, study_type, etc.) is preprocessed.

STEP 2 - Vectorization: Using PubMedBert (Our best-performing model), we compute vectors of 700 dimensions.

STEP 3 - Index Creation: Convert each vector into an index using **FAISS**.

STEP 4 - Similarity Search: Calculate the similarity between the query vector and reference vector (index) using cosine similarity to get the top k similar vectors.

Metric to calculate similarity between query vector (A) and reference vector (B)

$$\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|}$$

RESULTS

Metrics Reported:

1. **NER** (Named Entity Recognition): How well the model represents biomedical entities?
2. **SS** (Sentence Similarity): Can the model compute how similar are two biomedical sentences?`

Based on Biomedical Language Understanding and Reasoning Benchmark (BLURB)

Model Name	NER	SS
PubMedBERT	86.17	94.49

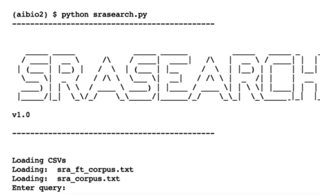
Based on our experiments with different language models, PubMedBert performs the best due to its vastness of biomedical vocabulary

Term	Type	BERT	SciBERT	PubMedBERT (Microsoft)
diabetes	disease	X	X	X
DNA	gene	X	X	X
RecA	gene			X
acetyl-transferase	gene			X
clonidine	drug			X

CONCLUSION

Our research shows how existing NLP and IR models can help redefine and significantly improve SRA search, providing greater flexibility to researchers and thus reducing time, complexity, and search errors.

Utilizing Neural Network based approaches, we show how we can enrich the NCBI SRA Search platform with contextual understanding. Using biomedical NLP models, we are able to encode biological texts into computable vectors that represent meaning and can therefore significantly reduce inaccurate search results.



Register interest to be the first one to experience SRASearch!

REFERENCES

[1] Zhu, Yuelin, et al. "SRAdb: query and use public next-generation sequencing data from within R." *BMC bioinformatics* 14.1 (2013): 1-4.

[2] Choudhary, Saket. "pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive." *F1000Research* 8 (2019).

[3] Leinonen, Rasko, et al. "The sequence read archive." *Nucleic acids research* 39.suppl_1 (2010): D19-D21.

[4] Gu, Yu, et al. "Domain-specific language model pretraining for biomedical natural language processing." *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1 (2021): 1-23.

[5] Wolf, Thomas, et al. "Transformers: State-of-the-art natural language processing." *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations.* 2020.

[5] Paszke, Adam, et al. "Pytorch: An imperative style, high-performance deep learning library." *Advances in neural information processing systems* 32 (2019).

ACKNOWLEDGEMENT

Funding Support: SJSU Level-Up Grant