Developing a semantic event search engine for biomedical events

Julio C. Rangel[1], Norio Kobayashi[1]

[1]RIKEN Information R&D and Strategy Headquarters 2-1 Hirosawa, Wako, Saitama, 351-0198 Japan

Abstract

We introduce Biomedical Event Finder (BEF), a semantic search engine that finds biomedical events (BE) from PubMed documents. BEF can be accessed via a web interface or a RESTful API. The simple retrieval of biomedical events from journal articles using keyword-based tools suffers from the issue of keyword mismatch; therefore, a key feature of our search engine is that it achieves semantic textual similarity via event embeddings to produce the most comparable results. We describe the primary components of our tool and our preliminary efforts

Keywords

semantic search, search engine, biomedical event, PubMed

Introduction

The scientific literature has a massive quantity of relational knowledge, which includes connections between biological entities such as proteins, medications, and symptoms. Researchers are increasingly interested in extracting meaningful, organized, short, and clear information about biomedical events to deal with the ever-increasing volume of publications. Biomedical processes are modeled using biomedical event structures. Generally, they consist of signal words, known as the event's trigger, and biological things, known as the event's arguments. The trigger determines the event type, which specifies the semantics (or roles) of its parameters. Typically, event triggers are nouns or verbs, such as oxidation, transcription, or reproduction, whereas biological substances are usually proper nouns, such as DNA, ATP, or lactase. The role theme defines the primary subject of attention in an event, whereas the role cause is frequently the event's facilitator or driver. Figure 1 depicts a typical event structure in the biomedical field.

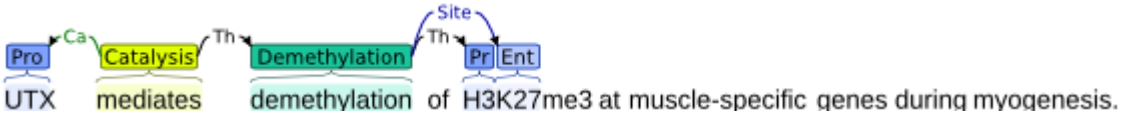

Figure 1: Example biomedical event

Even though the number of tools to extract biomedical events is growing each year [1],

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

juliocesar.rangelreyes@riken.jp (J. C. Rangel); norio.kobayashi@riken.jp (N. Kobayashi)

existing search engines are not taking advantage of the data produced by such tools. In this work, we first use DeepEventMine to create a dataset of events extracted from the PUDMED baseline, which is then leveraged to develop a Faiss index to allow semantic search. To the best of our knowledge, we are the first to attempt to construct a semantic searcher for biomedical events on the whole PUBMED database.

Extracting biomedical events

First, we extracted cancer genetics (CG) and infectious disease (ID) events from 34 million PubMed abstracts and titles using DeepEventMine (DEM). Because it generates over 1 TB of BRAT standoff files, it is impractical to use the DEM tool directly on all PubMed records.The standoff format is then omitted, and events are converted into a NetworkX graph and saved as smaller JSON files. In addition, we increased GPU data processing to increase DEM's speed.

Index creation

We generated a sentence-BERT embedding for each event graph by concatenating the names of each node. In addition, index storage is reduced by decreasing the embedding dimension with BERT-whitening [2]. The resulting embeddings are converted into a Faiss flat index. To allow search by year and event type (CG or ID), we created one index file for each year and event type.

Semantic Search

BEF relies on a semantic search BERT layer to compare the event query to the year and event type indexes. A query is embedded using Sentence-BERT and compared and rated against the event-type and year indexes.

Future work

Ongoing development is being done to enhance the performance of the engine as well as add and expand functionalities. We will add GENIA, epigenetics (EPI), pathway curation (PC), and MLEE to the list of biomedical events. To enhance the precision of the search, we intend to develop and evaluate a variety of event embedding methods. In addition, we intend to generate user-friendly visualizations for the events and normalize them according to ontologies.

References

[1] Q. Li, J. Li, J. Sheng, S. Cui, J. Wu, Y. Hei, H. Peng, S. Guo, L. Wang, A. Beheshti, P. S. Yu, A compact survey on event extraction: Approaches and applications 14 (2021) 1–21. URL: http://arxiv.org/abs/2107.02126.

[2] J. Su, J. Cao, W. Liu, Y. Ou, Whitening sentence representations for better semantics and faster retrieval (2021). URL: http://arxiv.org/abs/2103.15316.