Chan Zuckerberg Initiative

Building the Data Citation Corpus

Ana-Maria Istrate, Senior Research Scientist

Webinar • Feb 22, 2024

CZ Science

Supporting the science and technology that will make it possible to cure, prevent, or manage all diseases by the end of this century.

Open Science

Universal and immediate open sharing of all scientific knowledge, processes and outputs

We want to identify and democratize emerging and valuable **methods, tools**, and **datasets** and bring them to a broad and diverse set of scientists

so that they can come to meaningful conclusions faster

We create and share datasets on key research resources

GLOBAL BIODATA COALITION

CZI/GBC collaboration to surface biodata resources from full-text papers to build the Global Biodata Resource Inventory

DRYAD

CZ Software Mentions: A large dataset of software mentions in the biomedical literature

Dataset of software mentions from the biomedical literature (CC0)

Data Citation Corpus

Methods

Datasets

- Data aggregators (DataCite, Wikidata) have made it easier to discover datasets
- However, they don't have **comprehensive coverage**
  - Many domain specific repositories are not included
  - Not all datasets have DOIs
  - Majority of datasets are mentioned (not formally cited) in full-text of papers
  - We can use machine learning to extract these datasets from papers

extract datasets from the source

Datasets

GSE40279
https://identifiers.org/geo:GSE40279

GSE51032
https://identifiers.org/geo:GSE51032

https://doi.org/10.17632/RT6X6362YX.1

Dataset Accession Number IDs

Methylome data were downloaded from Hannum et al[5] and EPIC[26] (Gene Expression Omnibus, ***GSE40279*** and ***GSE51032***) and were processed alongside the methylation data generated from our sample.

https://doi.org/10.1001/jamanetworkopen.2020.15428

GSE40279
https://identifiers.org/geo:GSE40279

GSE51032
https://identifiers.org/geo:GSE51032

DOIs

Data associated with this study has been deposited at Mendeley Data under the accession number ***https://doi.org/10.17632/RT6X6362YX.1***.

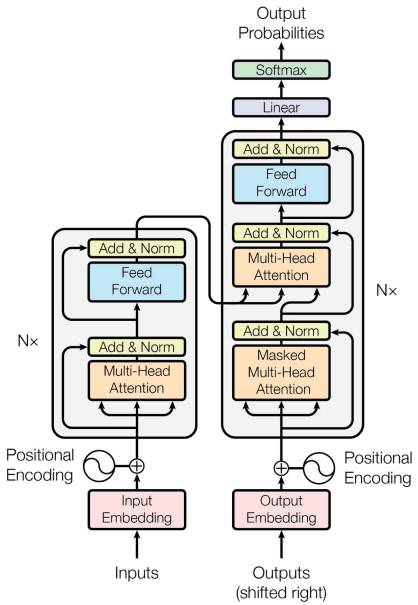
https://doi.org/10.1016/j.heliyon.2020.e05507

https://doi.org/10.17632/RT6X6362YX.1

How We Did It

Machine Learning Model

SciBERT-based Named Entity Recognition



Corpus

Europe PMC Full-Text 5M papers

Extraction

Retrieve dataset mentions and the repository they are linked to

The microarray data had been previously deposited at Gene Expression Omnibus (GEO) under accession number **GSE2603.**

prediction: B-DAT-GEO
dataset in GEO database

Linking

Link to a repository

https://identifiers.org/geo:GSE2603

Training Data

- subset of Europe PMC annotations, curated by our in-house team of biomedical curators
- 44 total repositories

Evaluation

- performed manually by our in-house biocuration team on a subset of 200 papers from EPMC that were not included in the training data

model	precision	recall	F1 score
SciBERT	0.741	0.980	0.844
GPT2	0.901	0.852	0.875

gen	biosample	rfam
pdb	dbgap	treefam
nct	emdb	empiar
geo	metagenomics	hgnc
refseq	interpro	rrid
uniprot	biostudies	efo
refsnp	cath	intact
bioproject	hipsci	go
ensembl	gisaid	uniparc
igsr	metabolights	rnacentral
pxd	ega	hpa
arrayexpress	reactome	biomodels
pfam	ebisc	orphadata
eudract	complexportal	
gca	chembl	

Repositories mined

Linking

links strings extracted by the ML model to URLs in identifiers.org

label Outputted by the NER model Will be B-DAT- + x, where x is one option below	**Linking Methodology** Where dataset is the extracted_word (or data mention)
arrayexpress	https://identifiers.org/arrayexpress:dataset
biomodels	https://identifiers.org/biomodels.db:dataset
bioproject	https://identifiers.org/bioproject:dataset
biosample	https://identifiers.org/biosample:dataset
..	

- We are validating the links by checking the URL responses. The final file only contains **URLs return a status_code of 200**

The microarray data had been previously deposited at Gene Expression Omnibus (GEO) under accession number **GSE2603.**

↓ NER Model

GSE2603

↓ Linking

https://identifiers.org/geo:GSE2603

Examples

Data Availability Coordinates and cryo-EM maps data have been deposited in Protein Data Bank (PDB) and Electron Microscopy Data Bank (EMDB) (PDB: 7RA3 , EMDB: EMD-24334) (PDB: 7RBT , EMDB: EMD-24401) (PDB: 7RGP , EMDB: EMD-24453) (PDB: 7RG9 , EMDB: EMD-24445).

https://www.pnas.org/doi/full/10.1073/pnas.2116506119

DAT-pdb

- 7RA3
- 7RBT
- 7RGP
- 7RG9

DAT-emdb

- EMD-24334
- EMD-24401
- EMD-24453
- EMD-24445

We aligned the 16S rRNA gene sequence of the strain NO (AJ132639) along with other members of the Arboriphilicus and Filiformis group (U41095, U82322, AB009827, AB026925, and AB065294), representative species of the genus Methanobrevibacter (U55233, U55240, U62533 and RDP-Mbb.rumina.

https://link.springer.com/article/10.1186/1471-2180-4-20

DAT-genbank

- AJ132639
- U41095,
- U82322
- AB009827
- AB026925
- U55233
- U55240
- U62533

False negatives

- Precision of the model is 0.741 and recall is 0.980 on our validation set, which means that we should be catching most mentions; however, some might still get missed
- Based on our analyses, will likely happen on long sentences
- during training, input is truncated to a maximum length
- some sentences, especially the ones in tables, are longer than this max_length, which means that input at the end of the sentences does not get seen by the model

DNA Deposition The following information was supplied regading the deposition of DNA sequences: Genomic data can be found at http://www.ncbi.nlm.nih.gov/bioproject/ , BioProject numbers: PRJNA209307 , PRJNA209312 , PRJNA209316 , PRJNA209319 , PRJNA209320 , PRJNA209333 , PRJNA209334 , PRJNA209340 , PRJNA209342 , PRJNA209343 , PRJNA209345 , PRJNA209347 , PRJNA209351 , PRJNA209352 , PRJNA209373 , PRJNA209375 , PRJNA209376 , PRJNA209465 , PRJNA209466 , PRJNA209468 , PRJNA209470 , PRJNA209476 , PRJNA209477 , PRJNA209479 , PRJNA209788 , PRJNA209483 , PRJNA209492 , PRJNA209493 , PRJNA209497 , PRJNA209498 , PRJNA209499 , PRJNA209501 , PRJNA209502 , PRJNA209504 , PRJNA209507 , PRJNA209508 , PRJNA209510 , PRJNA209512 , PRJNA209514 , PRJNA209517 , PRJNA209522 , PRJNA209596 , PRJNA209599 , PRJNA209616 , PRJNA209631 , PRJNA209633 , PRJNA209635 , PRJNA209637 , PRJNA209638 , PRJNA209639 , PRJNA209640 , PRJNA209642 , PRJNA209647 , PRJNA209648 , PRJNA209650 , PRJNA209654 , PRJNA209655 , PRJNA209659 , PRJNA209310 , PRJNA209310 , PRJNA209335 , PRJNA209641 .

→ DAT-bioproject

https://peerj.com/articles/806/

False positives

- Precision of the model is 0.741 and recall is 0.980 on our validation set, which means that we are catching most things at the cost of false positives
- This is a known tradeoff that most NLP models grapple with

Grant Numbers

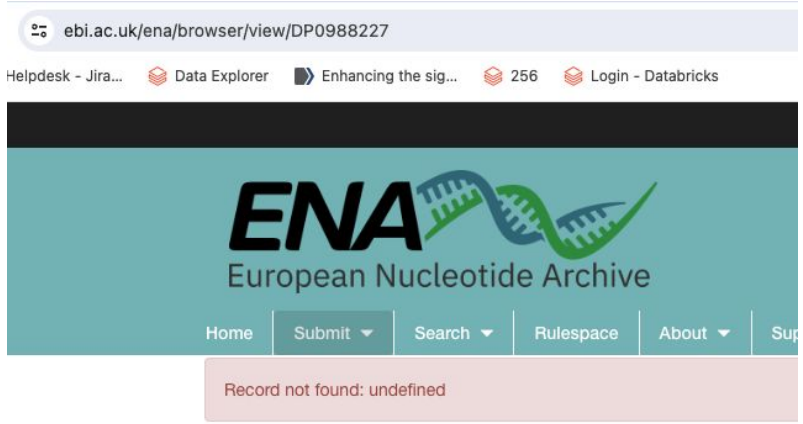
- some grant numbers will get identified as datasets, most often associated with GenBank entries
- likely to the accession numbers being similar to those in GenBank

Acknowledgements This work was supported by the Brain Foundation and Australian Myasthenic Association in New South Wales (to SWR), and Australian Research Council (DP0988227 to DAP).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5856814/

DAT-genbank

- DP0988227



we are cross-checking the status code of the URL returned, but that can return 200 even if no resource is found
mitigation: negotiate with the repo itself!

https://identifiers.org/ena.embl:DP0988227 comes back with status_code = 200

Other insights

~27%

dataset mentions

have links in full-text

~66%

datasets have the database name in full-text

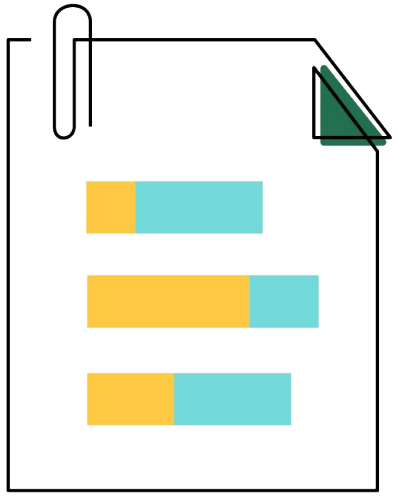
Takeaways and Future Directions

1. **Mining for accession number IDs is not trivial and not a solved task!**

2. **Mitigating false positives**
   a. Cross-checking that entries are valid in their predicted repository beyond checking the 200 status code returned by URL
   b. Use prompt engineering/ Large Language Model (e.g. GPT3.5/4)
      i. e.g: *You are a helpful open science assistant in charge of checking if a given data entry belongs to a data repository or not*
      ii. zero/few-shot
   c. Train a more powerful model (e.g. GPT3.5/4) for dataset extraction and compare performance with SciBERT model

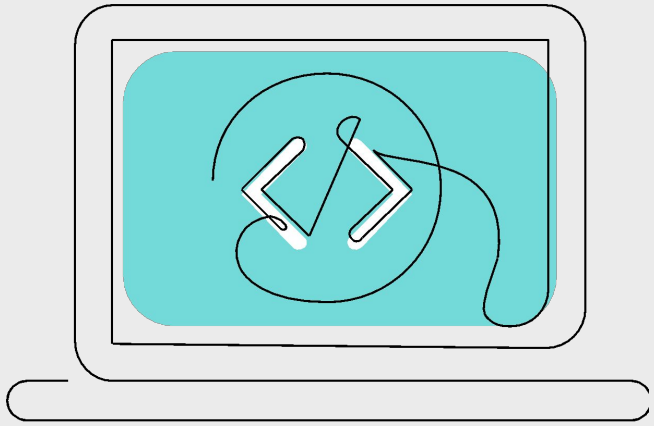
3. **Community developments/engagement!**

CZI Contribution to the Open Global Data Citation Corpus



seed datafile

dataset-paper links extracted with ML models from 5M Europe PMC Full-Text Papers



algorithms

new ML methodology in mining datasets from full-text papers

will be open-sourced, ETA end of March
Github + HuggingFace

Thank you!

CZI Science

@cziscience

https://medium.com/@cziscience

CZI Science Tech

https://tech.chanzuckerberg.com/scitech/

CZI Open Science

https://czi.co/OpenScience

Contributors

Ana-Maria Istrate
Senior Research Scientist, Science, model development

Michaela Torkar
Lead Curator, Science, curation

Fabrizio Castrotorres
Staff Software Engineer, Data Infrastructure, model deployment

with additional support from

Dario Taraborelli
Science Program Officer, Open Science

Jennifer Kennedy
Director, Software Engineering, Central Technology

Patricia Flores
Communications Manager, Science

Donghui Li
Senior Technical Program Manager, Science