Accelerating COVID-19 Research With Graph Mining and Transformer-based Learning

Ilya Tyagin, Ankit Kulshrestha, Justin Sybrandt, Krish Matta, Michael Shtutman & Ilya Safro

Abstract

We present automated transformer-based hypothesis generation systems AGATHA-C and AGATHA-GP for COVID-19 research. The systems are based on the graph mining and transformer models. They are massively validated and achieve high-quality predictions across multiple domains in fast computational time and are released to the broad scientific community to accelerate biomedical research. We also show that the systems are able to discover ongoing research findings such as the relationship between COVID-19 and oxytocin hormone.

Keywords: Natural Language Processing, Graph Mining, Deep Learning, Hypothesis Generation, Recommendation System, COVID-19.

Introduction

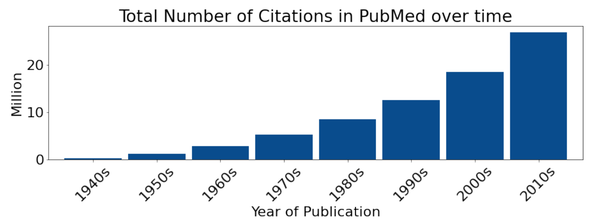

Figure 1: The pace of publishing scientific information is unprecedented and only grows over time.

To address the current COVID-19 data science challenge, we present AGATHA-C/GP hypothesis generation models and provide easy to use interface to broad scientific community. In addition, we evaluate the systems via candidate ranking using recent scientific publications.

Pipeline Summary

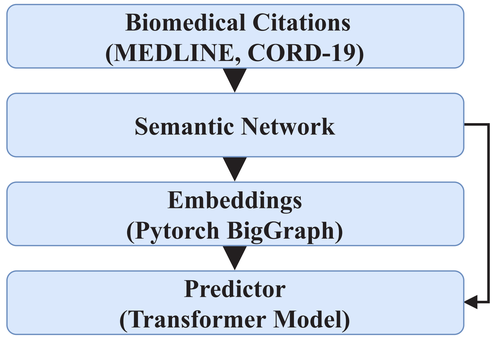

Figure 2: AGATHA pipeline diagram.

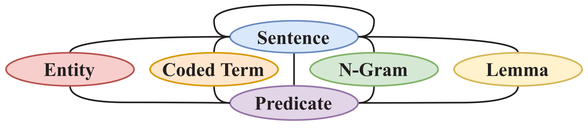

Figure 3: AGATHA multi-layered semantic network. Nodes on the schema represent different instance types and edges represent different types of connections.

Predicate Extraction

AGATHA-C pipeline utilizes NLM-developed tool SemRep to extract semantic predicates from the input citations.
To improve the scope of information retrieval, in our general purpose model AGATHA-GP we take advantage of deep learning OpenIE system provided by AllenNLP.

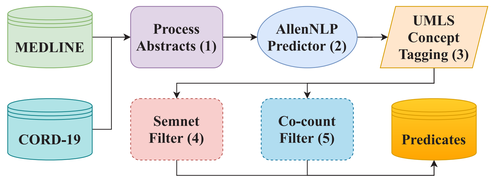

Figure 4: Predicate Extraction pipeline with Deep Learning-based Open IE system.

We couple it with UMLS Semantic Network and Co-count-based filtering to improve the quality of final predicates and connect them to UMLS Metathesaurus entries.

Massive Proactive System Validation

To present the results in terms of its value for urgent CORD-19-related hypothesis generation, we use a historical benchmark approach. For that we:

- Collect recent COVID-19-related citations published after the model training cut date
- Extract semantic predicates with the proposed techniques
- Filter out noise
- Categorize predicates to perform subdomain recommendation task.

Subdomain recommendation task:

- For each UMLS term collect its semantic type and group all extracted predicates by the term-pair criteria (combination of subject and object types)
- Identify the top-n most common term-pairs subdomains
- Construct the validation set from pairs belonging to these n subdomains
- For each positive pair randomly generate 10 domain-specific negative pairs.

	ROC AUC			PR AUC		
	O	C	GP	O	C	GP
dsyn:dsyn	0.83	0.88	0.88	0.35	0.41	0.42
phsu:dsyn	0.86	0.91	0.91	0.36	0.41	0.46
fndg:dsyn	0.85	0.93	0.92	0.40	0.54	0.54
dsyn:fndg	0.81	0.89	0.90	0.32	0.43	0.45
fndg:humn	0.81	0.89	0.89	0.37	0.45	0.48
dsyn:humn	0.77	0.84	0.85	0.27	0.33	0.37
topp:dsyn	0.87	0.92	0.92	0.38	0.52	0.50
orch:dsyn	0.87	0.89	0.88	0.34	0.45	0.45
geoa:spco	0.75	0.74	0.90	0.22	0.20	0.44
aapp:dsyn	0.86	0.93	0.92	0.39	0.44	0.49
Mean	0.83	0.88	**0.90**	0.34	0.42	**0.46**

Table 1: Classification quality metrics across recently popular COVID-19-related biomedical subdomains. Labels O, C and GP stand for AGATHA-O (baseline), C (COVID-19) and GP (General Purpose) models, respectively. Used abbreviations: *dsyn*: Disease or Syndrome; *topp*: Therapeutic or Preventive Procedure; *humn*: Human; *aapp*: Amino Acid, Peptide, or Protein; *phsu*: Pharmacologic Substance; *orch*: Organic Chemical; *spco*: Spatial Concept; *fndg*: Finding; *geoa*: Geographic Area.

Rediscovering Important COVID-19 Findings

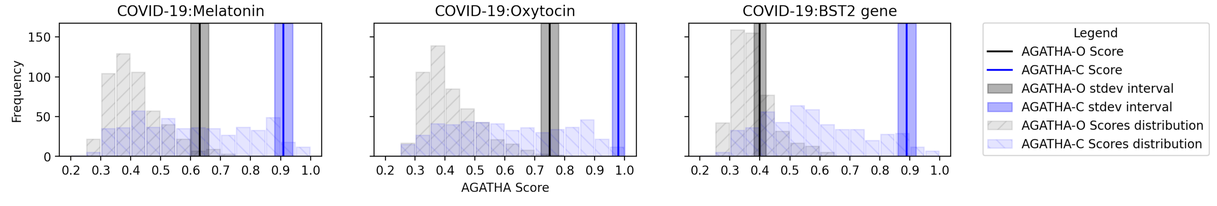

Figure 5: Score distributions in case study experiment. Presented scores are obtained with AGATHA-O (baseline) and AGATHA-C models. We test three COVID-19-related novel connections manually selected by the domain expert. These connections involve Melatonin and Oxytocin hormones and BST2 gene.

Training Data Sensitivity

We test how the inclusion of more recent training data affects AGATHA performance. Retraining the model using the proposed method and more recent data yields in slightly better scores in all basic metrics (Table 2).

Model	ROC AUC	PR AUC	AP.@10
C (Oct 2020)	0.88	0.49	0.67
C (June 2021)	0.90	0.55	0.77

Table 2: Comparison between AGATHA-C models trained with different cut dates.

References

[1] Ilya Tyagin, Ankit Kulshrestha, Justin Sybrandt, Krish Matta, Michael Shtutman, and Ilya Safro. Accelerating COVID-19 Research with Graph Mining and Transformer-based Learning, 2021.

[2] Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. *AGATHA: Automatic Graph Mining And Transformer Based Hypothesis Generation Approach*, page 2757–2764. ACM, New York, NY, USA, 2020.