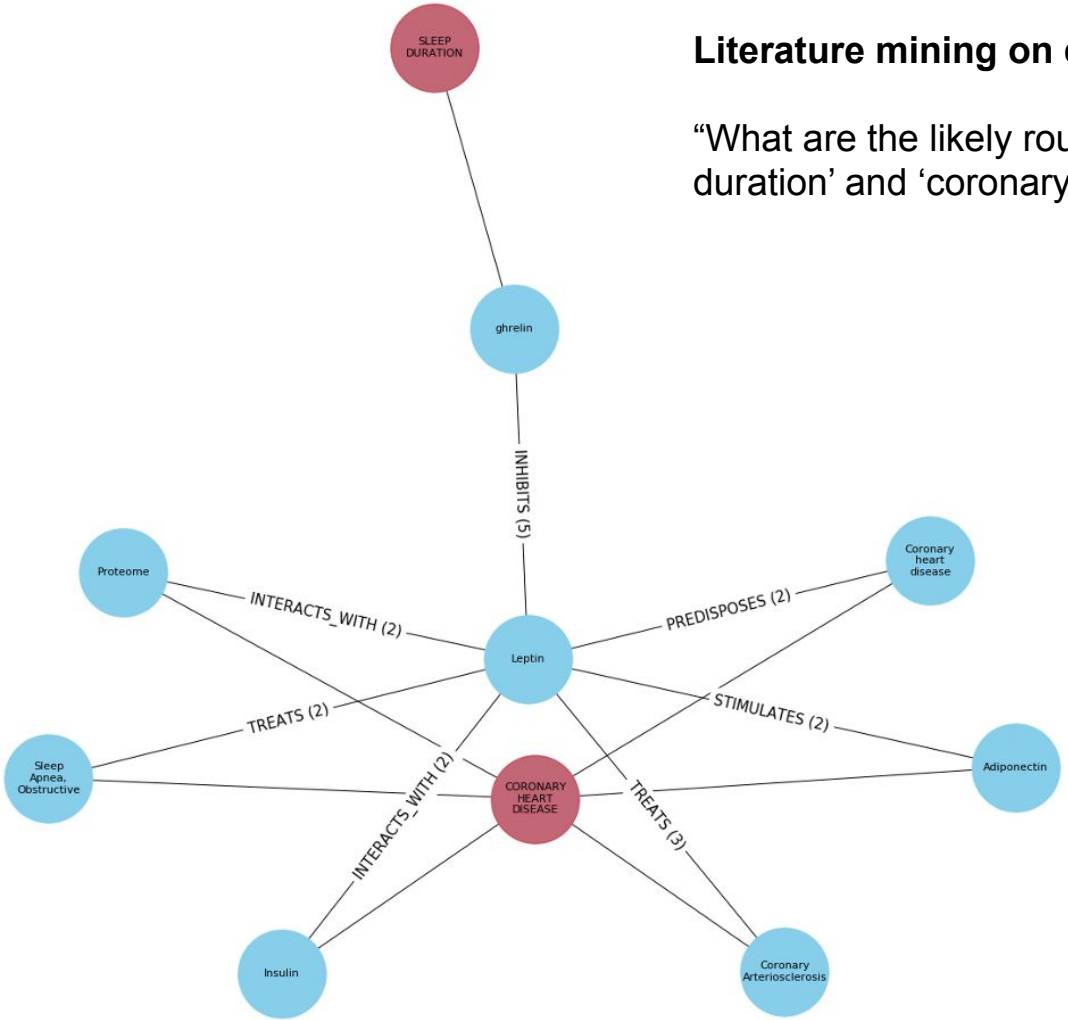
Automatic Graph-mining and Transformer based Hypothesis generation Approach (AGATHA)

Justin Sybrandt, Ilya Tyagin, Michael Shtutman, Ilya Safro

https://github.com/JSybrandt/agatha

Yi Liu, P4 virtual lunch paper talk, 2020-05-22



Literature mining on derived mechanisms

"What are the likely routes of associations between 'sleep duration' and 'coronary heart disease'?

Background: literature mining

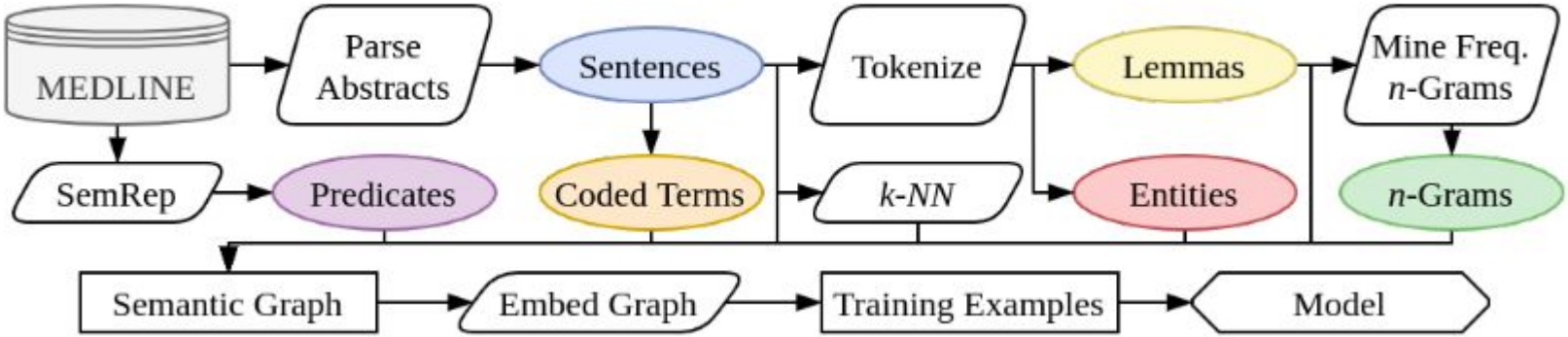
- MEDLINE: the database of biomedical and life sciences journal articles
- The Semantic Knowledge Representation project
  - SemRep: a software to extract “subject-predicate-object” triples
  - SemMedDB: a database of all SemRep triples of all PubMed articles
- Knowledge and knowledge graphs
  - Structured representation of heterogeneous entities
  - SemMedDB, MELODI; EpiGraphDB, Open Targets, Hetio, etc
- Modern NLP technologies
  - spaCy, python transformers, allennlp
  - Transformers (BERT)
  - Graph embeddings
  - biomedical language models: scipaCy, medspaCy, SciBERT, BioBERT

Background: MELODI (Elsworth et al., 2018)

- Overlapping terms
- An enrichment score statistic
- Architecture
  - Loads SemMedDB into a Neo4j graph
  - Queried via a Django web app
- MELODI-presto
  - Elasticsearch-based
  - subset of SemMedDB
  - Django-REST API + web app

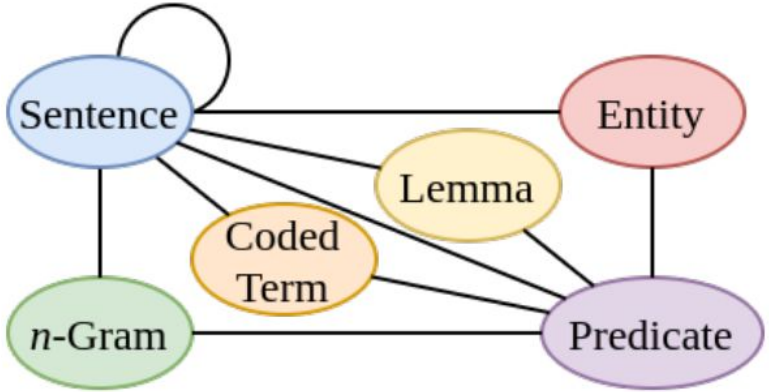
AGATHA: Architecture and building blocks

- A “deep-learning biomedical hypothesis generation system”
- Construction of a semantic graph
- Embed this graph
- Train a transformer-encoder model to learn this graph



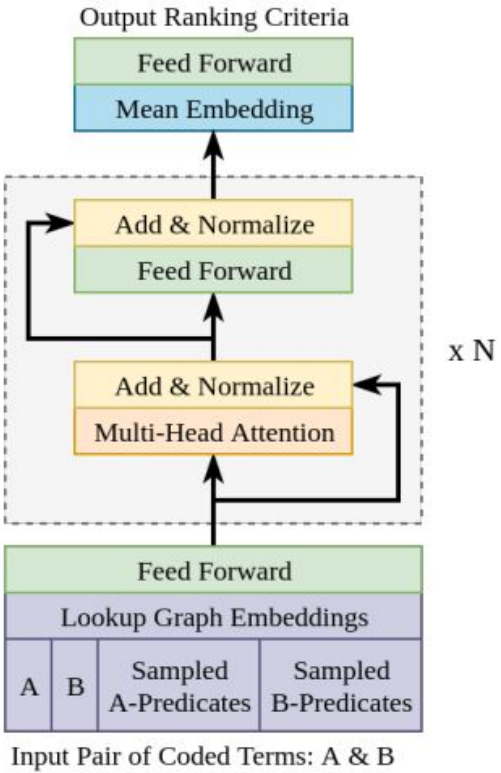
Block 1: the semantic graph

- Extract MEDLINE article metadata (PMID, version, title, abstract text, date, keywords) from 2015-01
- Split abstract into sentences. Use scispaCy to annotate parts-of-speech, dependency tags, and named entities. Extract n-grams on sentences.
- Use SciBERT (scibert-scivocab-uncased) to produce sentence embeddings, use FAISS for dimension reduction, then performs nearest neighbour on the embeddings
- Simple sentence-occurrence
- 184 million nodes, 12.3 billion edges



Block 2: Graph embedding and training

- Embedding
  - PyTorch-BigGraph: distributed
  - Two settings: 256-dim and 512-dim
- Training (transformer encoder)
  - Training data: a pre-2015 temporal holdout dataset of triples from SemMedDB
  - Task:
    - Rank SemRep predicates, minimize the ranking error
    - Positive / negative samples



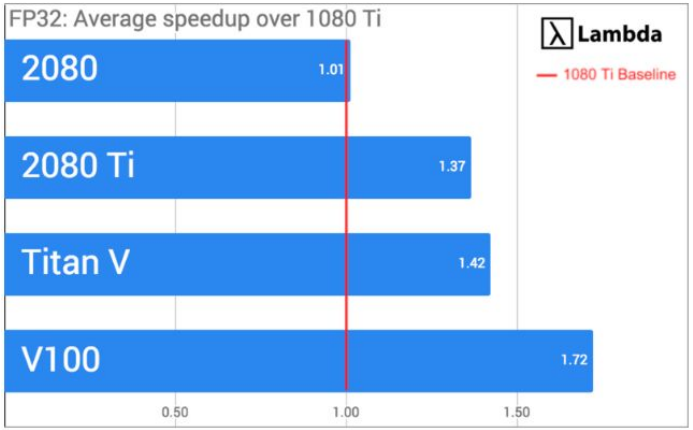
Validation

- Against Moliere (topic modelling based) (*Sybrandt et al., 2017, MOLIERE: Automatic biomedical hypothesis generation system*)
- Against Heuristic-based ranking using a ranking criteria from *Sybrandt et al., 2018, Large scale validation of hypothesis generation systems via candidate ranking*
- By subdomain (semantic type) recommendation
- Metrics: AUC ROC, AUC PR, top-k precision, average precision, mean average precision, mean-reciprocal rank

System Instance	ROC AUC	PR AUC
Moliere: Medline	0.718	0.820
Moliere: Full Text	0.795	0.778
AGATHA-256	0.826	0.895
AGATHA-512	**0.901**	**0.936**

Specs

- Graph embedding
  - A cluster of 20 nodes, each node with 24 CPU cores, 72h walltime
    https://www.palmetto.clemson.edu/palmetto/userguide_palmetto_overview.html
  - 256-dim embedding / 512-dim embedding
- Traning
  - LAMB optimizer
  - multi-GPU: 10 NVIDIA Tesla v100
- Techs
  - PyTorch-lightning, PyTorch-BigGraph, fire
  - scispaCy, SciBERT,
  - FAISS



https://lambdalabs.com/blog/best-gpu-tensorflow-2080-ti-vs-v100-vs-titan-v-vs-1080-ti-benchmark/