Graph Attentive Networks for Synonymy Prediction at Scale in the UMLS Metathesaurus

Goonmeet Bajaj
Summer 2021

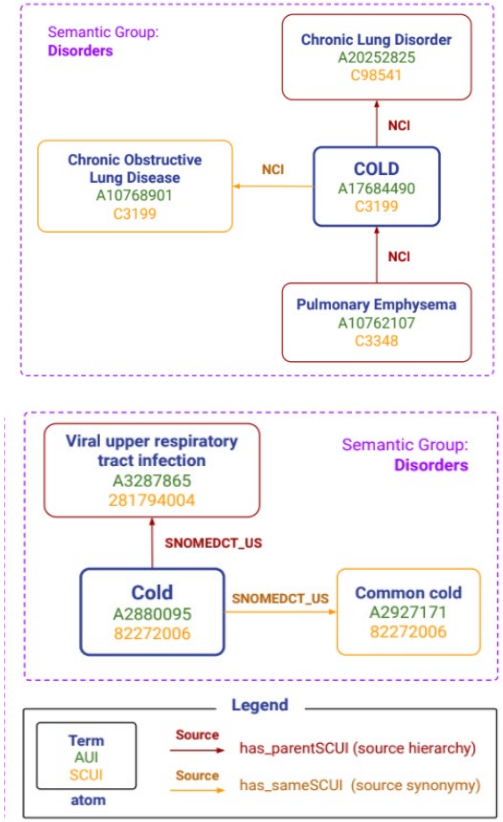
Mentors: Dr. Vinh Nguyen, Dr. Olivier Bodenreider

National Library of Medicine
Lister Hill National Center for Biomedical Communications

The Ohio State University

Motivation

- UMLS Metathesaurus integrates biomedical terms from various vocabularies
- Different vocabularies lead to different terms for similar concepts
- Current UMLS construction process: *tedious, error-prone, expensive*
- Our prior work [1, 2]:
  - Rule-based approximation of current construction process
  - LexLM: deep learning model that leverages lexical patterns
  - *ConLM: LexLM + knowledge graph embeddings*
- How can we leverage *contextual information*?
  - How does adding contextual information (i.e., semantic group, source synonymy, hierarchical information) affect disambiguation of terms?
  - Which graph-based models are suitable for synonymy prediction?



[1] Nguyen, V., Yip, H. Y., & Bodenreider, O. (2021, April). Biomedical Vocabulary Alignment at Scale in the UMLS Metathesaurus. In *Proceedings of the Web Conference 2021* (pp. 2672-2683).
[2] Yip H. Y., Nguyen, V., Sheth, A., & Bodenreider, O. Context-Enriched Learning Models for Aligning Biomedical Vocabularies in the UMLS Metathesaurus. Under submission

Objectives

1. ***Survey graph-based*** deep learning techniques for leveraging contextual information from UMLS.

2. ***Develop*** a ***novel, scalable, graph-based*** deep learning model using contextual information for synonymy prediction that ***outperforms LexLM & ConLM.***

Scalabilty Challenges

	DBP 15K*	Open Academic Graph**	UMLS
# of Nodes	55K to 105K	700 Million + (split across 3 graphs)	13 Million +
# of Training Pairs	153K to 279K	20K	118 Million +

Computational Limits:

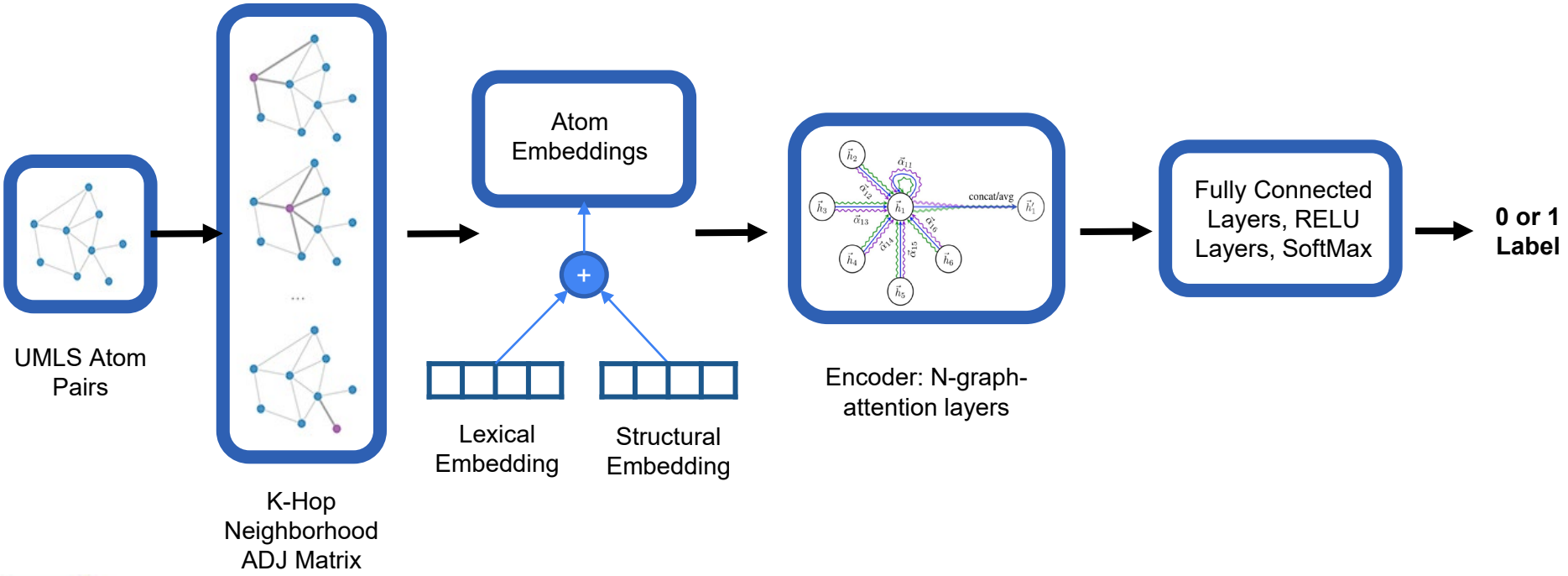
- *500+ GB to load training data (BioWulf limit: 373 GB on GPU node)*
- *7 Days for 1 Epoch on Single GPU*

*Sun, Z., Hu, W., & Li, C. (2017, October). Cross-lingual entity alignment via joint attribute-preserving embedding. In International Semantic Web Conference (pp. 628-644). Springer, Cham.
** Zhang, Fanjin, et al. "Oag: Toward linking large-scale heterogeneous entity graphs." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019.

Contributions

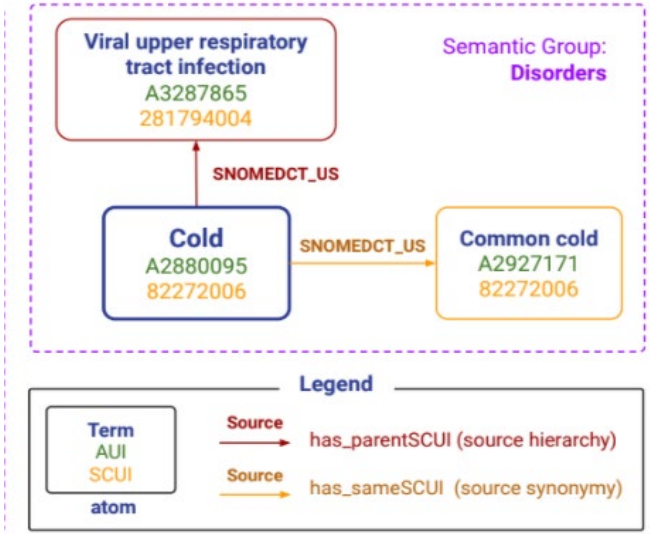
1. Implemented graph attention network (GAN) for synonymy predicion at scale
2. Evaluated and analyzed performance of GAN models
3. Identified shortcomings and areas of improvement of GAN

Graph Attention Network



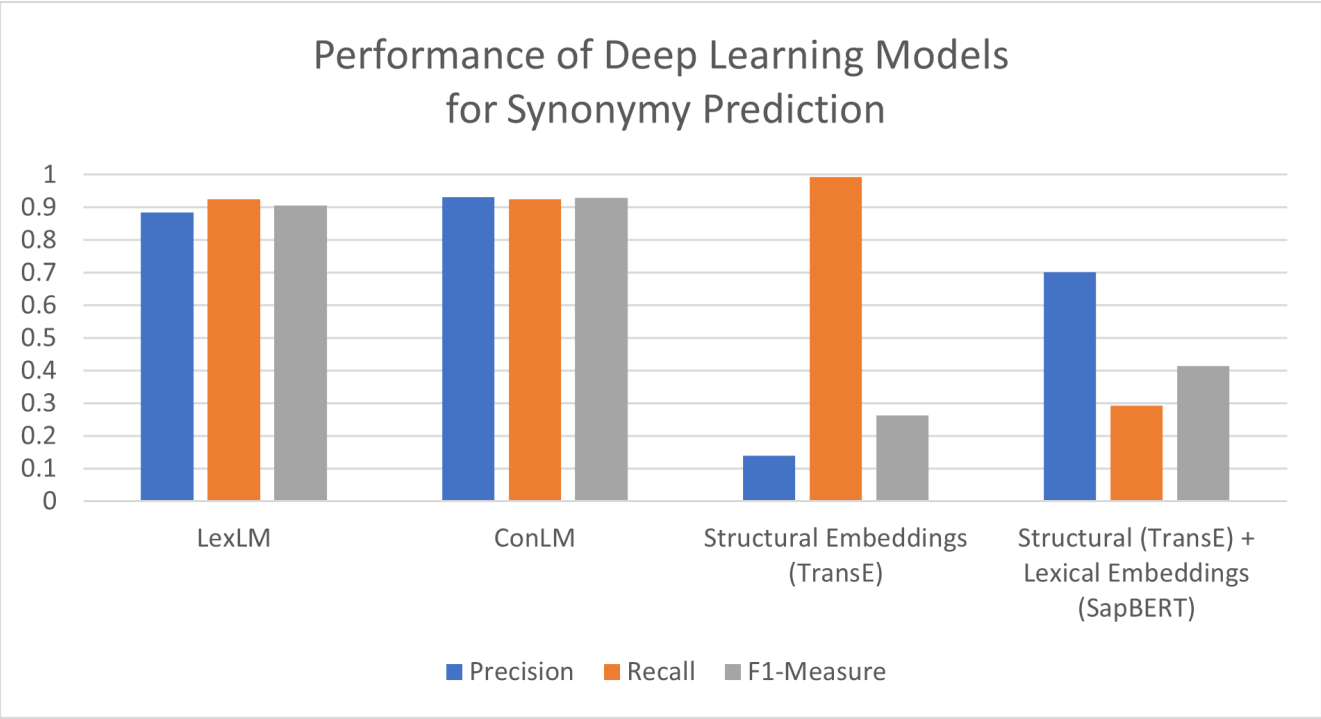
Model Components

- Graph structure
  - Directed / undirected graph
  - Homogeneous / heterogeneous graph (e.g., node type, edge type)
- Node / edge embeddings
  - Lexical embeddings: BioWordVec [1], SapBERT [2], UBERT, etc.
  - Graph structural embeddings: TransE [3], ComplEx [4], etc.



[1] Zhang, Yijia, et al. "BioWordVec, improving biomedical word embeddings with subword information and MeSH." Scientific data 6.1 (2019): 1-9.
[2] Liu, Fangyu, et al. "Self-alignment pretraining for biomedical entity representations." arXiv preprint arXiv:2010.11784 (2020).
[3] Bordes, A., et al. (2013). Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.
[4] Trouillon, Théo, et al. "Complex embeddings for simple link prediction." International conference on machine learning. PMLR, 2016.

Experimental Design & Results



Current progress

- Ongoing Work:
  - Training different model variants
  - Develop a new GAN model
  - Conduct qualitative analysis with different GAN models
- Future Work:
  - Develop novel graph embedding method
  - Explore heterogeneous graph transformer network

Acknowledgements

Dr. Bodenreider

Dr. Vinh Nguyen

&

My fellow interns: Thilini, Vishesh

Team: Joey Yip, Dr. Kin Wah Fung, Dr. Yuqing Mao