Pretraining Language Models for Synonymy Prediction at Scale in the UMLS Metathesaurus

Thilini Wijesiriwardene

Mentors:
Dr. Vinh Nguyen
Dr. Olivier Bodenreider

NIH National Library of Medicine
Lister Hill National Center for Biomedical Communications

AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

Motivation

- UMLS Metathesaurus integrates biomedical terms from various vocabularies
- Current UMLS construction process: tedious, error-prone, expensive
- Our prior work*:
  - LexLM: a deep learning model leveraging lexical patterns
  - Rule-based approximation of current construction process
- Can more recent techniques in Deep Learning and NLP perform better in UMLS Metathesaurus construction?

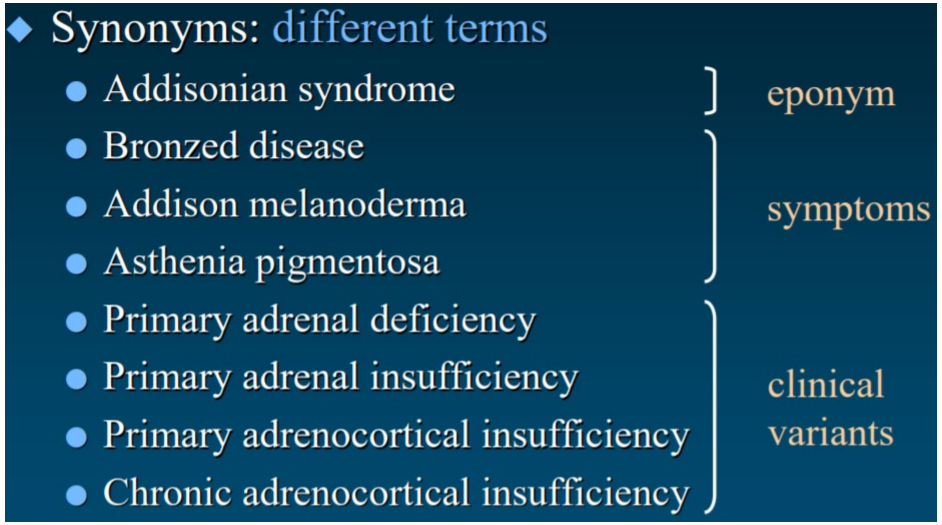
Objectives

Develop UBERT, a BERT based language model pretrained on UMLS data and Synonymy Prediction task that can provide state-of-the-art performance on Synonymy Prediction

*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)

Synonymy Prediction Task*

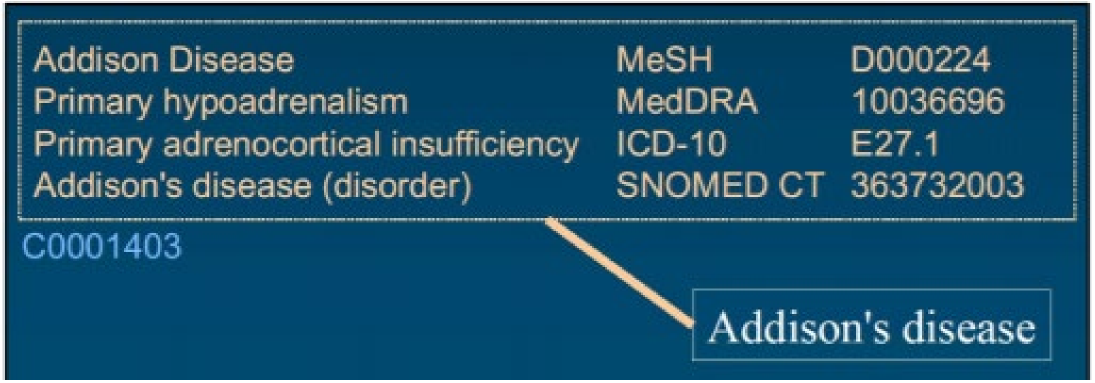
“atom” – Single term form a source vocabulary



- Synonyms: different terms

Addisonian syndrome	eponym
Bronzed disease	symptoms
Addison melanoderma	
Asthenia pigmentosa	
Primary adrenal deficiency	clinical variants
Primary adrenal insufficiency	
Primary adrenocortical insufficiency	
Chronic adrenocortical insufficiency	

atoms with same meaning are grouped in to one concept identified by a Concept Unique Identifier (CUI)



Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

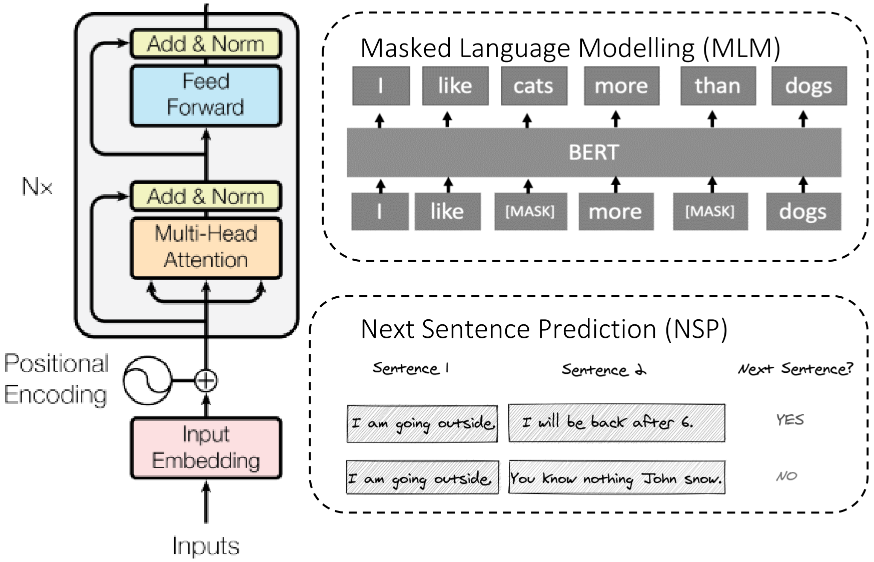
C0001403

Addison's disease

*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).

BERT (Bidirectional Encoder Representations from Transformers)*

Artificial neural network-based language model, designed to provide meaning for a word by using its surrounding context.

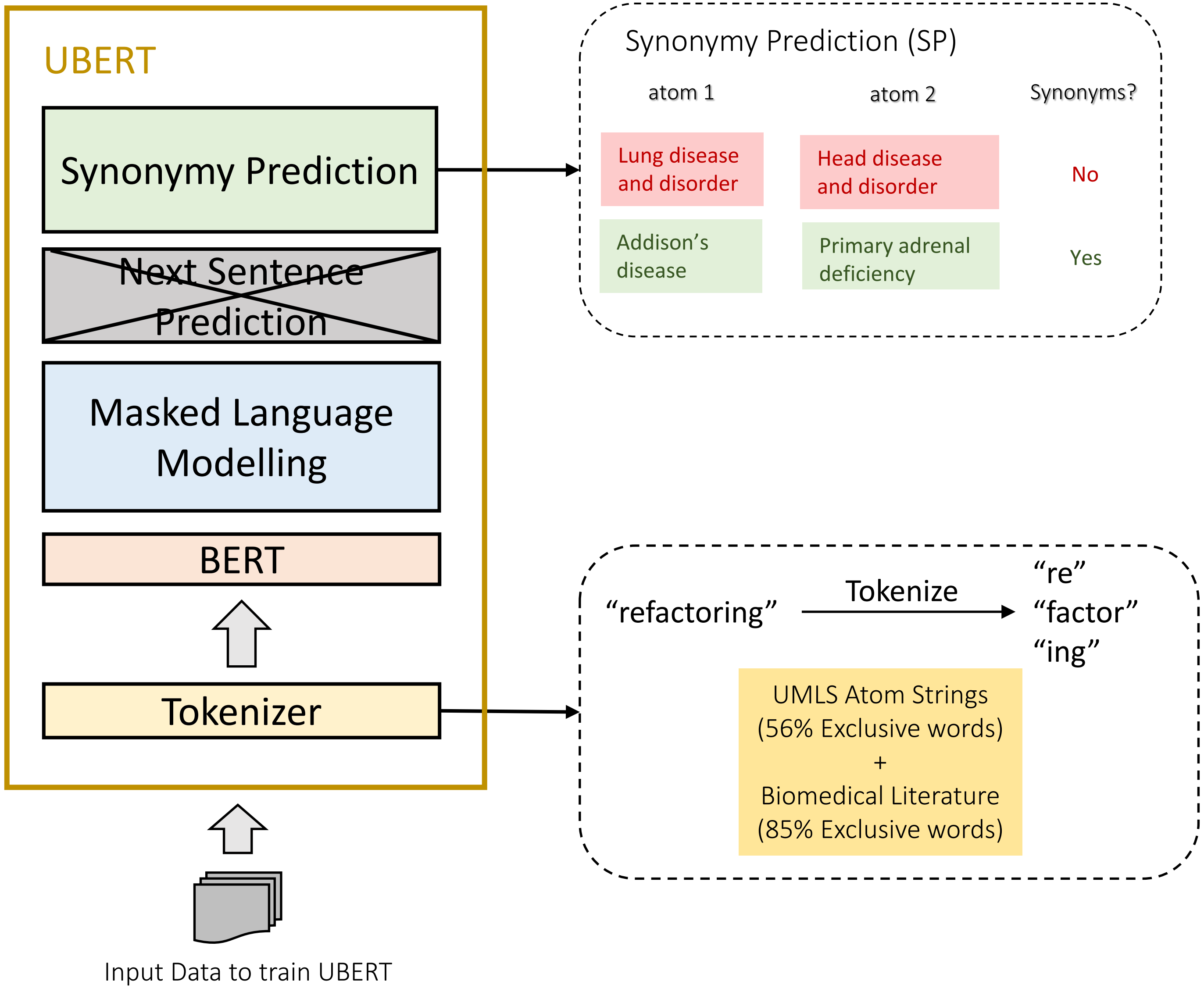


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs
Masked Language Modelling (MLM)
I like cats more than dogs
BERT
I like [MASK] more [MASK] dogs
Next Sentence Prediction (NSP)

Sentence 1	Sentence 2	Next Sentence?
I am going outside.	I will be back after 6.	YES
I am going outside.	You know nothing John snow.	NO

BERT Architecture

UBERT Architecture



UBERT
Synonymy Prediction
~~Next Sentence Prediction~~
Masked Language Modelling
BERT
Tokenizer
Input Data to train UBERT
Synonymy Prediction (SP)

atom 1	atom 2	Synonyms?
Lung disease and disorder	Head disease and disorder	No
Addison's disease	Primary adrenal deficiency	Yes

"refactoring"
Tokenize
"re"
"factor"
"ing"
UMLS Atom Strings (56% Exclusive words)
+
Biomedical Literature (85% Exclusive words)

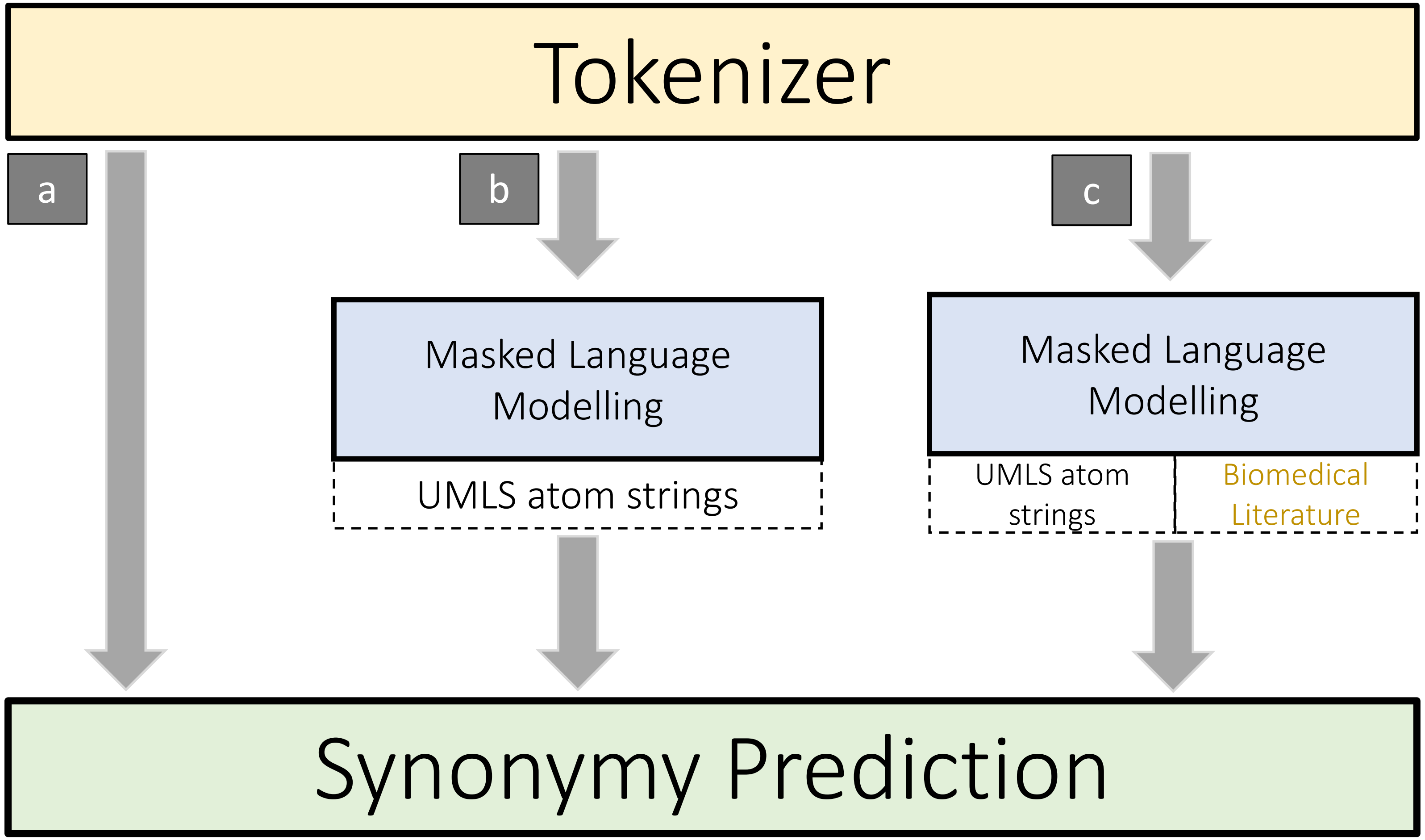
*Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1).

Datasets for Training & Testing UBERT

- Train tokenizer and Masked Language Modelling task
  - PubMed Abstracts → ~ 4.5 billion words
  - PubMed Central Full Texts → ~ 13.5 billion words
- Synonymy Prediction task :
  - Training → 118 million pairs
    - Positive (synonymous) pairs
    - *Negative (non-synonymous) pairs - varying degrees of lexical similarity
  - Testing → 170 million pairs

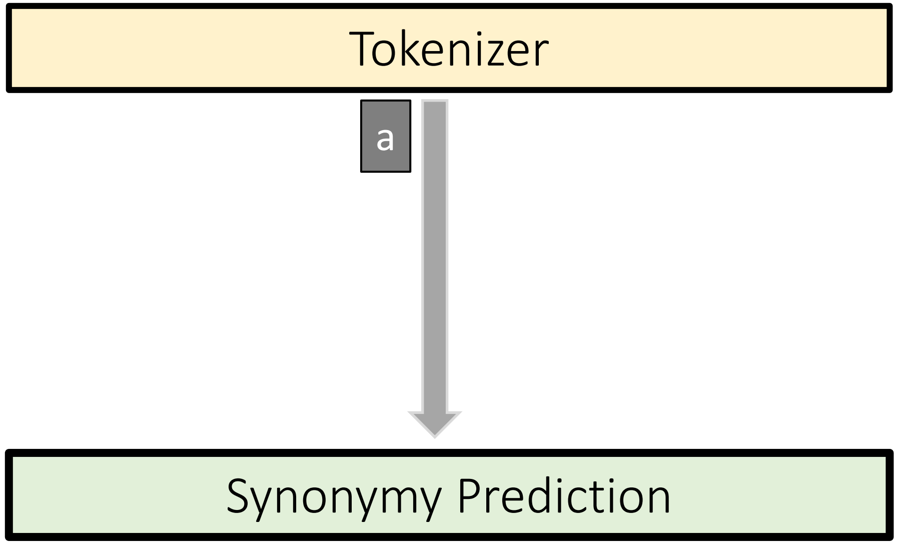
*Nguyen, V., Yip, H. Y., & Bodenreider, O. (2021, April). Biomedical Vocabulary Alignment at Scale in the UMLS Metathesaurus. In *Proceedings of the Web Conference 2021* (pp. 2672-2683).

Training Setup



National Library of Medicine
Lister Hill National Center for Biomedical Communications

Initial Results



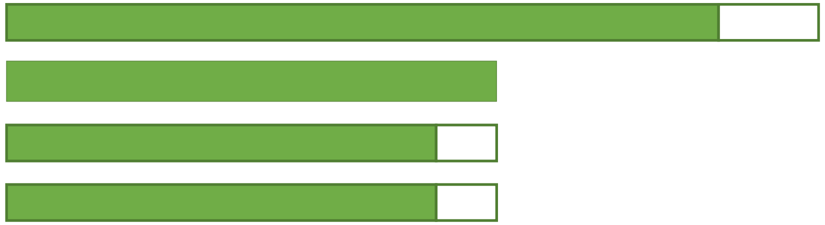
Metric	RBA	LexLM	UBERT
F1	0.7651	0.9061	**0.9974**
Accuracy	0.9863	0.9938	**0.9950**
Precision	0.8631	0.8875	**0.9991**
Recall	0.6871	0.9254	**0.9957**

Training UBERT for a single epoch takes around ~3 hours on 16 Nvidia V100X GPUs.

Internship Progress

Develop UBERT, a BERT based language model pretrained on UMLS data and Synonymy Prediction task that can provide state-of-the-art performance on

1. Synonymy prediction
    - Variant a
    - Variant b
    - Variant c



Conclusions and Future Work

- UBERT is a potential candidate for UMLS Metathesaurus construction
- Look for more efficient architectures for Synonymy Prediction task that leverage atoms’ contexts when predicting synonymy.
- Evaluating performance of UBERT in BioNLP tasks such as biomedical Named Entity Recognition, biomedical Relations Extraction, etc.

Acknowledgements

Dr. Olivier Bodenreider

Dr. Vinh Nguyen

Goonmeet

Vishesh

Joey

Dr. Kin Wah Fung

Dr. Yuqing Mao

Thank You
Questions?

References

1. Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. 3 Nucleic acids research, 32(suppl_1), D267-D270.
2. https://medium.com/genei-technology/richer-sentence-embeddings-using-sentence-bert-part-i-ce1d9e0b1343
3. https://amitness.com/2020/02/albert-visual-summary/
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1).
5. 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)
6. https://huggingface.co/transformers/model_doc/bert.html