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

Thilini Wijesiriwardene

Introduction

The Unified Medical Language System (UMLS) Metathesaurus is a large biomedical thesaurus which integrates concepts with similar meaning from nearly 200 different source vocabularies. UMLS Metathesaurus construction process aims to cluster synonymous terms from these varied source vocabularies under the same concept referred to as Concept Unique Identifier (CUI). The terms from the source vocabularies are the building blocks of the UMLS Metathesaurus and are identified as atoms or unique atom identifiers (AUI). Human editors are involved in the process of clustering synonymous AUIs to similar concepts. Due to the overwhelming size of the Metathesaurus with over 4 million concepts the process of integration is tedious, error-prone and time consuming. Transfer learning allow models to learn from certain tasks and use the learned representations in other similar tasks. Language model pretraining is a popular transfer learning approach in natural language processing (NLP) where the Deep Learning (DL) models are trained through self-supervised learning tasks on huge datasets. Inspired by the recent successes of pretrained bidirectional encoder representations from transformers (BERT) and related models in biomedical NLP tasks, we propose UBERT: a BERT based language model pretrained on UMLS and PubMed data and evaluate its performance on identifying synonymous terms in the UMLS Metathesaurus to aid the UMLS Metathesaurus construction process.

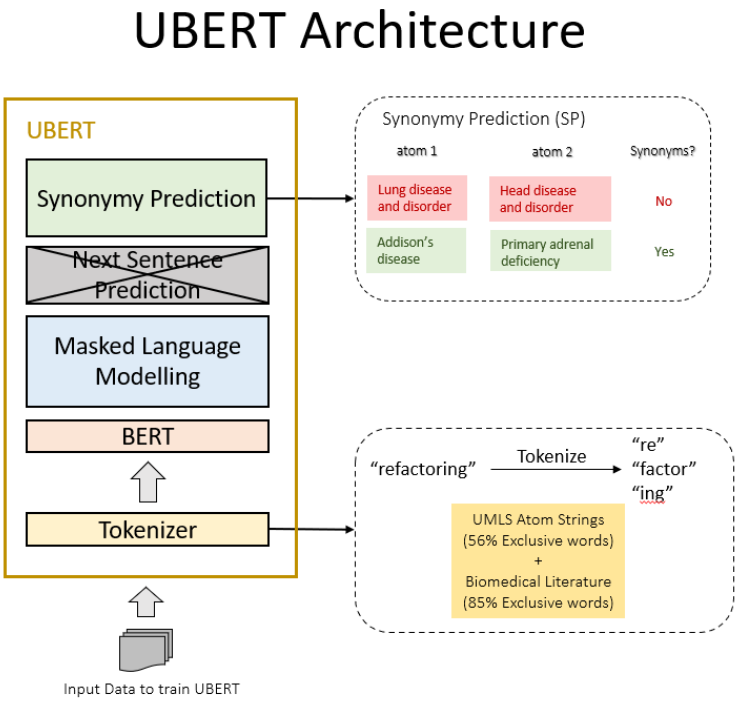

Figure 1: UBERT Architecture

Objectives

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

Methods

In this work we are utilizing the same BERT architecture pretraining mechanism from the original BERT paper [1] (huggingface implementation [2]) and perform pretraining using two datasets.

1) Synonymy dataset [3]
2) PubMed abstracts + PMC full articles (biomedical literature) [4]

BERT is pretrained using two tasks self-supervised training tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In our work, instead of using NSP we use Synonymy Prediction (SP) as a supervised training task where we train the model with pairs of AUI strings annotated as synonyms or non-synonyms (Figure 1).

There are three main components in training UBERT. First component is tokenization which is being used to tokenize the input data for MLM and SP tasks. We adopt the wordpiece tokenization approach by used by original BERT, which mitigated the concern for out-of-vocabulary words since wordpiece tokenization can represent any new word by corresponding subwords (e.g. refactoring → re, factor, #ing) Upon analysis of overlapping words from both the corpora we identified that 56% of the unique words in UMLS are exclusive to UMLS and 85% of the unique words found in biomedical literature are exclusive to biomedical literature. Therefore, we chose to and construct a wordpiece tokens vocabulary from the UMLS and biomedical literature corpora instead of reusing generic vocabulary provided by the pretrained BERT model.

The tokenized input is then sent through 3 variants of UBERT for pretraining based on the task used for pretraining and the dataset used for training. The variants and the tasks and data used for pretraining are listed below.

1) UBERT trained with synonymy dataset and SP task

   This variant of BERT is trained only using SP task in a supervised training manner. The tokenized AUI string pairs are sent through the BERT architecture and trained for 50 epochs.

2) UBERT trained with MLM task and then SP task
   a. UBERT trained with MLM task (UMLS data only) then SP task
      In this variant the MLM is pretrained with AUIs from the UMLS and further pretrained on the SP task.
3) UBERT trained with MLM task and then SP task
   a. UBERT trained with MLM task (UMLS data only) then SP task
      In this variant the MLM is pretrained with AUIs from the UMLS and biomedical literature, then further pretrained on the SP task.

Results

The first variant of UBERT has completed its training and validation runs and the results are as follows for the testing dataset: Accuracy = 0.99364, F1 = 0.99671, Precision = 0.99897, Recall = 0.99447, AUC = 0.94435. The training and testing of the second and third variants are ongoing.

Conclusion

We see that UBERT variant trained only on SP task outperforms previous Siamese architecture. Training UBERT was resource intensive as all the other transformer-based models (training and validation of UBERT variant 1 for 50 epochs took 8 days with 16 v100x GPUs). Further analysis can be done to identify more efficient DL architectures that can leverage the contextual information of the UMLS to predict synonymy.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.

[2] https://huggingface.co/transformers/model_doc/bert.html

[3] Nguyen, Vinh, Hong Yung Yip, and Olivier Bodenreider. "Biomedical Vocabulary Alignment at Scale in the UMLS Metathesaurus." Proceedings of the Web Conference 2021. 2021.

[4] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.