SPL-BERT: a deep bidirectional transformer language model using Structure Product Labeling (SPL) documents for approved drug products

Saad Haider, Joshua Xu, Weida Tong and Leihong Wu

Division of Bioinformatics and Biostatistics, NCTR, FDA

INTRODUCTION

Extracting and mining information from text-based documents using natural language processing has advanced significantly in recent years and is being applied to diverse biomedical documents including publications, electronic health records and product adverse event reports. Deep learning-based language models trained on domain specific texts can perform various language tasks such as name entity recognition (NER), question answering (Q&A), relation extraction (RE), etc.

The FDA uses Structured Product Labeling (SPL) standard for disseminating information of regulated products including ~130,000 drug products. Particularly, SPLs contain specific information related to drug safety and efficacy that current generic language models were not well trained on. Hence, a specific language model was developed by learning the SPL documents.

HIGHLIGHTS

Objective:

Development of a domain specific language representation model with structured product labeling (SPL) resources at FDA, to serve/help the FDA regulatory scientific researches such as drug labeling review or data analysis.

AIMs:

- ❑ Collection and extraction of information from SPL documentations.
- ❑ Development of a language model with appropriate model architectures and fine-tuning them with SPL data.
- ❑ Validation of the developed model on several drug-labeling specific tasks, such as Q&A, Adverse Reaction Alerts, summarizations, etc.

Data:

A total of ~6980 drugs with unique trade names contains boxed warning in the FDALabel database. These boxed warning texts were extracted and were used to pretrain the SPL-BERT after initializing the weights from BioBERT.

Approaches:

- ❑ Data preprocessing: generate analysis-ready datasets with SPL documentations for language modeling. Texts were extracted from XML files for each SPL document.
- ❑ Modeling: develop a specific BERT model based on pre-trained model BioBERT by further training on SPL datasets.
- ❑ Validation: evaluate the results from the new developed model and compare its performance to the current pre-trained models.

METHODS

Pipeline/Framework

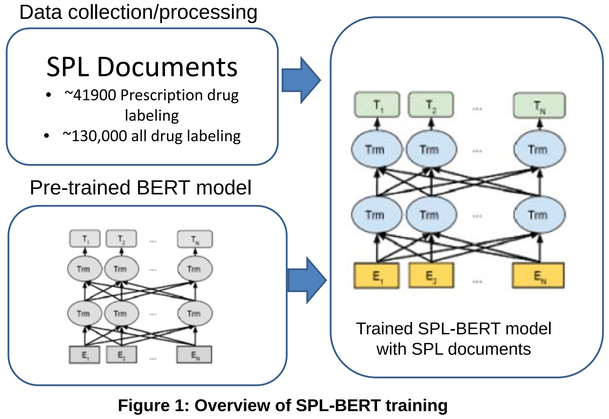


Figure 1: Overview of SPL-BERT training

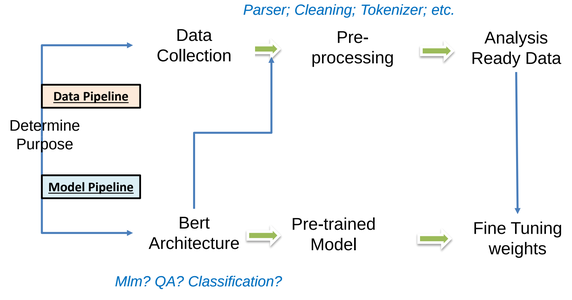


Figure 2: Pipeline for SPL-BERT data preprocessing and fine-tuning

Summary of SPL Data from FDALabel

Table 1 shows the number of sentences and words extracted from each section of the 41900 SPL documents. Table 2 shows list of text corpora used in SPL-BERT and its pre-trained checkpoint BioBERT. BioBERT was also build on pretrained weights for original BERT.

Table 1: SPL Drug Labeling document summary

SPL Document Section	Lines/Sentences	Words/Tokens
All	18,873,490 (18.87m)	294,200,471 (294.2m)
Boxed Warnings	134,559 (0.13m)	2183051 (2.18m)
Adverse Reactions	1,758,729 (1.76m)	27,848,588 (27.85m)
Indications and Usage	324,553 (0.32m)	6,277,040 (6.27m)
Contraindications	181,402 (0.18m)	2,598,821 (2.60 m)
Drug Interactions	539,746 (0.54m)	7,585,967 (7.59m)
Dosage and Administrations	1,134,449 (1.13m)	18,394,113 (18.39m)

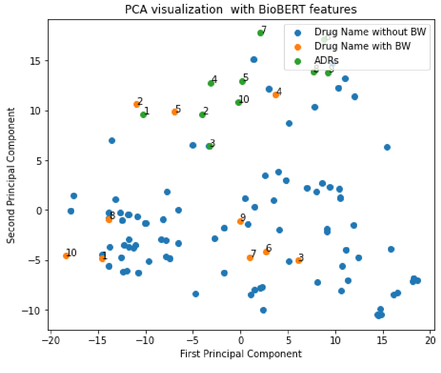
Table 2: List of text corpora for SPL-BERT

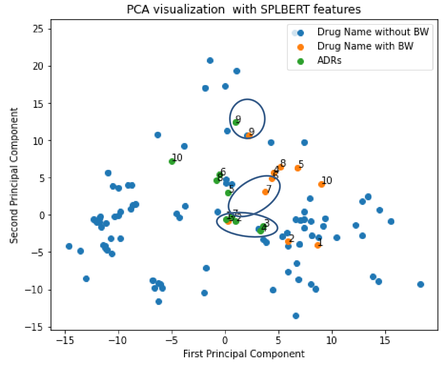
Corpus	Number of Words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical
SPL Drug-Labeling	0.3B	Drug-labeling

Table 3: Comparison between text corpora for BERT, BioBERT and SPL-BERT

Model	Corpus Combination
BERT-base	Wiki + Books
BioBERT Large	Wiki + Books + PubMed + PMC
SPL-BERT	Wiki + Books + PubMed + PMC + SPL

RESULTS





Label	Drug Name	Adverse Drug Reaction (ADR)
1	XYWAV	respiratory depression
2	VYXEOS	Hemorrhage
3	Xultophy 100/3.6	thyroid C-cell tumors
4	APTENSIO XR	abuse and dependence.
5	6% HETASTARCH IN 0.9% SODIUM CHLORIDE	Renal replacement therapy
6	EPZICOM	hypersensitivity reaction
7	ABACAVIR, LAMIVUDINE AND ZIDOVUDINE	Hematologic toxicity
8	Triumeq	fatal hypersensitivity reaction
9	TRIZIVIR	anemia
10	ZIAGEN	Lactic acidosis

Figure 3: Comparison of PCA visualization of features extracted from BioBERT and SPL-BERT. The features were extracted for a maximum sequence length of 128 and batch size of 8 for both BioBERT and SPL-BERT. The table on the right contains the description of the drug name and ADR labels seen in the PCA figures. It is seen that SPL-BERT grouped the ADRs much closer together than BioBERT. We expect this as SPL-BERT was fine-tuned with boxeds warning texts which contains the listed ADRs. In the PCA visualization for SPL-BERT, Xultophy (drug#3) is grouped near thyroid C-cell tumors (ADR#3). Similarly, ABACAVIR, LAMIVUDINE AND ZIDOVUDINE (drug#7) is grouped near its corresponding ADR hematologic toxicity (ADR#7). Also TRIZIVIR (drug#9) is grouped with anemia (ADR#9). We did not see such grouping in BioBERT visualization. Average distance between the green and orange points in 16.4 for BioBERT and 6.6 for SPL-BERT – indicating a better recognized association between drug names and their corresponding ADRs.

TECHNICAL DETAILS

HPC Node and GPU	1 node with 1 Nvidia Tesla V100
Number of GPU	1
Corpus Size	29 MB (2.18 million words)
Maximum Sequence Length	128
Steps	3500
Time to finish training	3 hours

Table 4: The computing resources used inside NCTR HPC.

One single node consisting of 1 Nvidia Tesla V100 GPU was used to pretrain SPL-BERT using only boxed warning texts from the 6890 drugs with unique trade names. The total time to train was around 3 hours. The next step is to train SPL-BERT with the full text from all sections of ~41900 drug labels.

DISCUSSION

There are existing deep learning-based language models which are trained on large publicly-available text corpus such as the English Wikipedia, BookCorpus, PubMed Abstracts, etc. No such model exists which focuses only on domain specific corpus such as drug SPL labeling. Drug labeling contains unique information about drugs that were not included in training datasets for those existing language models. The new SPL-BERT model trained with SPL boxed warning maps the word embedding features more accurately than BioBERT. Our model shows that pre-training BERT model on labeling documents helps it understand labeling specific texts. The output of the fully trained SPL-BERT using all texts will work as a novel resource for scientific research inside FDA for easier application of regulatory practices.

REFERENCES

1. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2019, issue 1367-4803 DOI: 10.1093/bioinformatics/btz682

2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,* arXiv:1810.04805

3. FDALabel, Full Text Search of Drug Product Labeling. https://nctr-crs.fda.gov/fdalabel/ui/search