Representation Learning On Drug Labels for Predicting Drug-Drug Interaction (DDI)

Qingyuan Liu[1]

[1]Toyota Technological Institute at Chicago

qingyliu@uchicago.edu

Abstract

It has become a popular topic in recent years to emphasize two or more chemicals administered concurrently. There are over 190,000 cases reported a drug affects another drug's efficacy. To target this issue, an algorithm predicting whether two drugs may impact interactions on one another is necessary. Here, we propose a method, by extracting information from drug labels, mainly from its Active Ingredient, Active Moiety, Indication and Usage, Contraindication, and Adverse Reactions, the drug's information is embedded to a unique 5*768 matrix and fed as inputs for prediction. A total of four-sentence embedding methods all based on Bidirectional Encoder Representations from Transformers (BERT) are applied. The prediction is not very accurate but does suggesting the embedding works to some extent and the drug pair who have interactions is different than those who do not by analyzing through the embedded space.

Index Terms: drug-drug interaction, sentence embedding, transformer, representation learning, BERT

1. Introduction

It is a well-known fact that you should not take alcohol while on sleeping pills as combining these increases the risk of overdosage on sedating effects. The interaction between pharmaceutical drugs is much greater due to its high specificity and high concentration nature.

To prevent harmful drug-drug interaction (DDI) from happening, it is strongly suggested by medical professionals that you should carefully and thoroughly read drug labels (a pamphlet came with the drug) and drug facts printed on the backside of the package before taking any medication. Due to its complicated nature of pharmaceutical terminology, understanding from an unprofessional can be imperfect and results in misled information.

We propose to build an algorithm to replace medical professionals by extract information from drug labels and use this information to predict drug-drug interactions. Language processing algorithm BERT[1] and its derivative SciBERT[2], BioBERT[3], and self-trained PharmBERT are applied to the drug label text to generate embedded vectors of the information. The generated vectors are later served as features for predicting DDI.

2. Methods

2.1. Datasets

2.1.1. Drug label dataset

The full prescription drug labels are downloaded from the DailyMed government server on March 8th, 2021. (https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-drug-labels.cfm) The dataset consists of 42850 prescription drug labels in XML format. Among those 42850 drug labels, the majority are duplicated drug labels, i.e the label for the same medicine but at a different time or different version. After redundancy reduction and only kept the copy the most up-to-date, a total of 6550 drug labels remained. These 6550 drug labels served as the main text corpus for this project.

The dataset is cleaned by removing all the garbled text such as those unsuccessfully encoded figure and table. The dataset is further manually corrected on those wrong text due to human errors.

2.1.2. Drug-drug interaction dataset

The DDI dataset is curated by drugbank.com specialist and is kindly referred from Ryu's paper supplementary data[4]. This dataset provides 192287 reported DDI pairs. Due to the fact this project only considering prescription drugs, not over-the-counter(OTC) drugs, with additional mapping issues between drugbank.com's unique drug ID, a total of 26441 DDI are able to be retrieved from 510 different medications. Among these 510 different medications, it is assumed if a drug pair is not denoted within those 26441 DDI, there will be no DDI between this drug pair. With this assumption, a total of 103354 (510*509/2-26441) no-DDI pair is inferred. The combined 26441 DDI and 103354 no-DDI would help on examining the performance.

2.2. PharmBERT training

PharmBERT is a Bidirectional Encoder Representations from Transformers fine tuned on a field-specific text corpus. This idea is admitted for other BERT variants as well. SciBERT[2] retrained the model on full papers from the corpus of semanticscholar.org with its 1.14 millions papers corpus size and 3.1 billion tokens.[2] BioBERT[3] tuned the model with PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) which consists of 4.5 billions tokens and 13.5 billions tokens respectively. Additional field-specific corpus largely improved performance for field-specific language processing. PharmBERT, pre-trained from SciBERT parameters, further tuned the model from the drug label dataset consists of a total of 6550 drug labels.

The training strategy is to train the model through mask filling. Text corpus is randomly masked at random word position. The language model is trained to predict what the masked word is. Each line of the corpus would count as a unique training sample and each training sample is limited to less than 128 words.

2.3. Drug labels embedding

Each drug label may or may not contain all the sections in FDA's drug label templates. However, nearly all of the drug

labels have the following 5 sections: Indication and Usage, Dosage and Administration, Contraindication, Warnings and Precautions, Adverse Reactions. Due to the lengthy nature and short time of this project, Warnings and Precautions did not include in the drug labels text embedding. Neither did the Dosage and Administration section as it has less to do with drug characteristics. The rest three sections (Indication and Usage, Contraindication, Adverse Reactions) are embedded as vectors to represent the drug's property.

Two additional section from drug labels is captured for feature generation as well. The first is the active ingredient of the drug, the second is the active moiety of the drug. Both section only include one phrase. These two sections extracted with the aim to capture any chemical structure information of the drug. All five section's text is fed to the sentence embedding methods.

One hyper-parameter of the embedding model needs to be tuned, the maximum word count for single sentence/single input. This hyper-parameter is mostly dominated by the hardware limit. Longer sentence takes significantly longer time to compute and will result in not enough memory. However, a longer sentence could capture more global information. Based on the word count distribution 1 maximum single sentence word count is set to the limit of 512. To speed up the embedding process, for each section, only the first 1000 words are considered. For sections that require two or more sentences, the output for each sentence from that section will be averaged.

The embedding of each drug's label will have dimension: 5 (extracted sections from drug labels)*768 (Embedding vector's dimension for each section)

2.4. Distance between drug labels

Squared Euclidean distance metric between each dimension is used to calculate the differences between two embedded drug labels. Let us denote two embedded drug labels, A, B, with A^1, A^2, B^1, B^2 as the embedding vector for active ingredient and active moiety; A^3, A^4,A^5, B^3, B^4, B^5 as the embedding vector for Indication and Usage, Contraindication and Adverse Reaction sections; $Euc^2()$ as the function to calculate squared euclidean distance between two vectors

$$Dist_1(A,B) = \begin{Vmatrix} Euc^2(A^1,B^1) & Euc^2(A^1,B^2) \\ Euc^2(A^2,B^1) & Euc^2(A^2,B^2) \end{Vmatrix} \quad (1)$$

$$Dist_2(A,B) = \begin{Vmatrix} Euc^2(A^3,B^3) & \dots & Euc^2(A^3,B^5) \\ Euc^2(A^4,B^3) & \dots & Euc^2(A^4,B^5) \\ Euc^2(A^5,B^3) & \dots & Euc^2(A^5,B^5) \end{Vmatrix} \quad (2)$$

$$Dist(A,B) = Dist_1(A,B) + Dist_2(A,B) \quad (3)$$

2.5. Drug-drug interaction prediction

2.5.1. Unsupervised method

It is detailed reasoned that the difference between two drug labels would be smaller if two drugs have DDI. The reasoning is as follows: Assume drug A, drug B both lower the blood pressure and drug C raise the blood pressure, drug D lower cholesterol. From drug labels text embedding, since both drug A and drug B lower the blood pressure, the "Lower the blood pressure" term would appear on both the drug labels, cause the two embedded labels to have the least difference. Between drug A and drug C, as they both affect blood pressure, "blood pressure" is on both drug labels, therefore, they are relatively similar. Between drug A and drug D, given that they have different functions, their text embedding would be much different. Therefore, the distance between two embedded drug labels will be shorter if they have DDI.

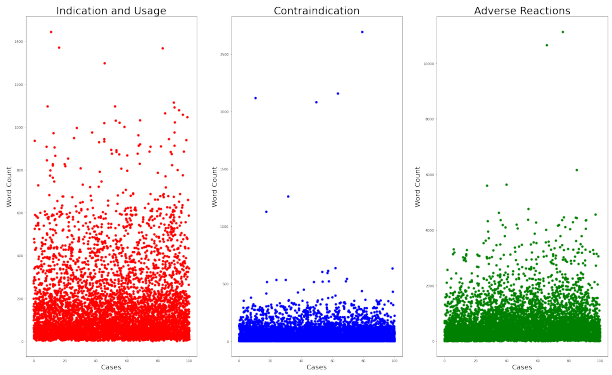

Figure 1: *Word count for Indication and Usage, Contraindication and Adverse Reactions.*

2.5.2. Supervised method

Given the known 26441 DDI drug pairs and 103354 no-DDI drug pairs, a fully connected binary classifier with two hidden layers of 1024 nodes and 256 nodes respectively is trained. The input of this binary classifier is the two embedded drug labels concatenated with dimension 10*768.

The training set is modified such that the model will be trained with a 1:1 encounter ratio of DDI drug pair and no-DDI drug pair to make sure the model encounter enough DDI data.

3. Results

3.1. PharmBERT mask filling

Two mask filling examples is shown in Figure 2. The PharmBERT did much better than SciBERT did. For the first sample, PharmBERT predicts correctly while Scibert does not. For the second example, the correct token is the 4th one, but does show up in the token list, SciBERT does not. PharmBERT is better at predicting high-frequency word tokens than low-frequency word tokens. This happened to other cases as well (Not shown). Overall, with additional corpus training, PharmBERT performed significantly BETTER than SciBERT.

3.2. Distance between drug pairs

Based on the DDI dataset, no-DDI dataset, and the distance calculation method 3, the distance between all the drug pairs can be calculated. It is plotted on Figure 3. A total of 4 embedding methods are tested. The results match the expectation that the DDI drug pair have a shorter distance(less different) than the no-DDI drug pair. It is observed for all four embedding methods. (at low distance region, more red points can be seen). If by using the distance of each drug pair as a matric and a better embedding model would have DDI drug pair's distance much less than the no-DDI drug pair's distance, PharmBERT performed best, followed by SciBERT, BERT, BioBERT. The average distance for both groups (DDI and no-DDI group) from each embedding methods is marked on Table 1. From the table, if use the difference of distance between DDI drug pair and no-DDI drug pair as a metric, PharmBERT performed significantly better than all other three embedding methods, with percentage

average difference of 4.74.

3.3. Drug-drug interaction prediction

3.3.1. Unsupervised Learning

Since by assumption, the DDI drug pair should have a lower distance than no-DDI drug pair, a distance cutoff is used to predict whether a drug pair have DDI or not. if the distance between a drug pair is less than a cutoff, it is noted as DDI drug pair. Based on different cutoff, the precision and recall is plotted for PharmBert and SciBERT methods, two of which performed best in embedding. The precison vs. recall plot is shown in Figure 4. From the precision vs. recall curve, PharBERT performed better than BioBert by producing higher precision when recall stays the same.

3.3.2. supervised Learning

Surprisingly, the fully connected binary classifier cannot correctly classify the datasets at all. The model either predicts all the data as DDI pair or no-DDI pair. The detailed reason is unknown. When monitoring the prediction output, as it should output a probability of two classes, the predicted probability of both classes is 0.

Table 1: *Distance average of different embedding methods*

Methods	**DDI Avg**
no-DDI Avg	**Diff%**
PharmBert	845.32
887.31	4.74
SciBERT	652.92
681.45	4.18
Bert	705.82
732.75	3.69
BioBERT	152.88
157.12	2.69

4. Discussion

Retraining BERT on the specific dataset or field-dataset will almost always produce a better result. SciBERT produced a better result than regular BERT. However, BioBERT, unforeseen, produced the worst result among all four BERT variants. Additionally, BioBERT embedding is the slowest and most RAM-required embedding model. The slow speed and large memory requirement may due to the large vocabulary token size BioBERT inherited. The detailed reason BioBERT performed the worst is unknown. Whether fine-tuning the BioBERT using drug label corpus would produce a better result is unsure. More experiments are needed for tunning different BERT models using a drug label corpus.

For determining the hyperparameter of the embedding models, the maximum word count per sentence is set a limit to 512. And each section only takes the first 1000 words. This hyperparameter should work for the Indication and Usage, and Contraindication section because from Figure 1, the word count for most of the Indication and Usage section and Contraindication section is smaller than the sentence limit, and word count limit, all the information can be fed to the embedding model. However, for the Adverse Reaction section, this is not enough. It is distinctly clear a lot of samples have Adverse Reaction's word count larger than 1000. Moreover, the Adverse Reaction section is probably the most important section among all the drug labels as a lot of DDI is due to the unaware side effects. Not knowing enough of the adverse reactions of the drugs would certainly limit the performance for classification.

For ease of conducting this project, all the embedded vectors for each sentence are averaged. This step probably lost a lot of essential local and global information in the text. A better pooling method is required or train a pooling method by measuring the later stage performance (DDI classification).

The distance function between two drug labels is calculated in two-part. Due to the different nature of the active ingredient, active moiety as they are mostly composed of one phrase instead of long sentences, finding the distance of embedded vectors between Active Ingredient, Active Moiety and Indication and Usage, Contraindication and Adverse reactions are illogical. And Active Ingredient and Active moiety sections are mostly for its chemical information, the rest three sections are for its drug description information.

It is unclear why the fully connected network is unable to classify the DDI pair. One hypothesis is that the gradient has vanished through backpropagation.

5. Conclusions and Future Direction

PharmBERT performed best among all BERT-derived methods tested on the drug labels dataset. The classification of this project is weak. A better classification model is required to better predict DDI drug pairs.

Future work will include setting up a different classifier that allows embedding of each sentence is used as a seperate input to maximumly gather information through the data. Another plan is to build an end-to-end model by using DDI prediction accuracy to affect embedding from drug labels.

6. Acknowledgements

Prof. Livescu and Freda Shi for reviewing and giving advice for the project.

7. References

[1] "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv.org*, no. https://arxiv.org/abs/1810.04805.

[2] "Scibert: A pretrained language model for scientific text," *arXiv.org*, no. https://arxiv.org/abs/1903.10676.

[3] Jinhyuk, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *OUP Academic, Oxford University Press*, no. https://academic.oup.com/bioinformatics/article/36/4/1234/5566506, 2019.

[4] Ryu and J. Y. et al, "Deep learning improves prediction of drug–drug and drug–food interactions." *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, p. doi:10.1073/pnas.1803294115, 2018.

```
#Diarrhea may be an early symptom of incomplete intestinal obstruction,
#especially in patients with ileostomy or colostomy.
fill_mask("[MASK] may be an early symptom of incomplete intestinal obstruction,\
especially in patients with ileostomy or colostomy.")
```

```
[{'score': 0.7008383870124817,
  'sequence': 'diarrhea may be an early symptom of incomplete intestinal obstruction
  'token': 16557,
  'token_str': 'diarrhea'},
 {'score': 0.15508608520030975,
  'sequence': 'vomiting may be an early symptom of incomplete intestinal obstruction
  'token': 18644,
  'token_str': 'vomiting'},
 {'score': 0.05971681699156761,
  'sequence': 'constipation may be an early symptom of incomplete intestinal obstruc
  'token': 29795,
  'token_str': 'constipation'},
 {'score': 0.022686924785375595,
  'sequence': 'nausea may be an early symptom of incomplete intestinal obstruction,
  'token': 18816,
  'token_str': 'nausea'},
 {'score': 0.015483860857784748,
  'sequence': 'fever may be an early symptom of incomplete intestinal obstruction, e
  'token': 10551,
  'token_str': 'fever'}]
```

```
#Diarrhea may be an early symptom of incomplete intestinal obstruction,
#especially in patients with ileostomy or colostomy.
fill_mask_regular("[MASK] may be an early symptom of incomplete intestinal obstr
especially in patients with ileostomy or colostomy.")
```

```
[{'score': 0.38035503029823303,
  'sequence': 'it may be an early symptom of incomplete intestinal obstruction, esp
  'token': 256,
  'token_str': 'it'},
 {'score': 0.1818888932466507,
  'sequence': 'this may be an early symptom of incomplete intestinal obstruction, e
  'token': 238,
  'token_str': 'this'},
 {'score': 0.08637246489524841,
  'sequence': 'dic may be an early symptom of incomplete intestinal obstruction, es
  'token': 10756,
  'token_str': 'dic'},
 {'score': 0.03182947263121605,
  'sequence': 'cd may be an early symptom of incomplete intestinal obstruction, esp
  'token': 1389,
  'token_str': 'cd'},
 {'score': 0.014126553200185299,
  'sequence': 'regurgitation may be an early symptom of incomplete intestinal obstr
  'token': 26717,
  'token_str': 'regurgitation'}]
```

```
#AYGESTIN (norethindrone acetate tablets USP) is indicated for the treatment of\
#secondary amenorrhea, [endometriosis], and abnormal uterine bleeding due to \
#hormonal imbalance in the absence of organic pathology, such as submucous
#fibroids or uterine cancer.

fill_mask("AYGESTIN (norethindrone acetate tablets USP) is indicated for the \
treatment of secondary amenorrhea, [MASK], and abnormal uterine bleeding due to\
 hormonal imbalance in the absence of organic pathology, such as submucous \
 fibroids or uterine cancer.")
```

```
[{'score': 0.23511268198490143,
  'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the t
  'token': 22890,
  'token_str': 'lactation'},
 {'score': 0.09662080556154251,
  'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the t
  'token': 17441,
  'token_str': 'abortion'},
 {'score': 0.082520011239528656,
  'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the t
  'token': 28441,
  'token_str': 'ovulation'},
 {'score': 0.055117227137088776,
  'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the t
  'token': 24730,
  'token_str': 'endometriosis'},
 {'score': 0.04050794243812561,
  'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the t
  'token': 21906,
  'token_str': 'infertility'}]
```

```
#AYGESTIN (norethindrone acetate tablets USP) is indicated for the treatment of\
#secondary amenorrhea, [endometriosis], and abnormal uterine bleeding due to \
#hormonal imbalance in the absence of organic pathology, such as submucous
#fibroids or uterine cancer.

fill_mask_regular("AYGESTIN (norethindrone acetate tablets USP) is indicated for
treatment of secondary amenorrhea, [MASK], and abnormal uterine bleeding due to\
 hormonal imbalance in the absence of organic pathology, such as submucous \
 fibroids or uterine cancer.")
```

```
{'score': 0.14431177079677582,
 'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the
 'token': 5564,
 'token_str': 'pregnancy'},
{'score': 0.10275698453187943,
 'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the
 'token': 2675,
 'token_str': 'pain'},
{'score': 0.08248776197433472,
 'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the
 'token': 28395,
 'token_str': 'hysterectomy'},
{'score': 0.039769694209098816,
 'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the
 'token': 5352,
 'token_str': 'infections'},
{'score': 0.03630823642015457,
 'sequence': 'aygestin ( norethindrone acetate tablets usp ) is indicated for the
 'token': 21906,
 'token_str': 'infertility'}]
```

Figure 2: *Two random examples of filling masks. fill_mask function is from PharmBERT and fill_mask_regular function is from SciBERT. The green text/comment is the correct sentence. In both cases, PharmBERT performed better than SciBERT. For the first sample, PharmBERT predicts correctly while Scibert*

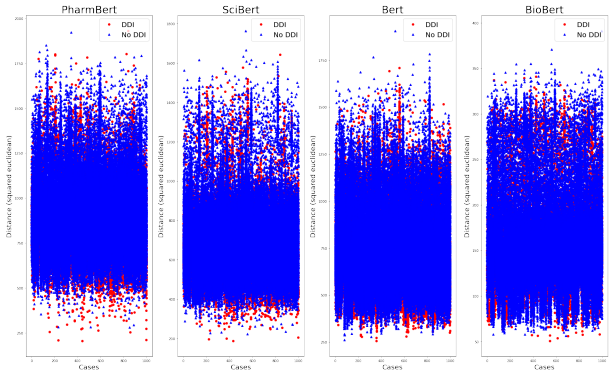

Figure 3: *A total of 4 embedding methods are tested. Distance between each drug pair for different embedding methods is illustrated. The results match the expectation that DDI drug pairs have shorter distances (less different) than no-DDI drug pairs. It is observed for all four embedding methods. (pay attention to the region that distance is low, more red points can be seen). If by using the distance difference between DDI pair and no-DDI pair, PharmBERT performed best, followed by SciBERT, BERT, BioBERT.*

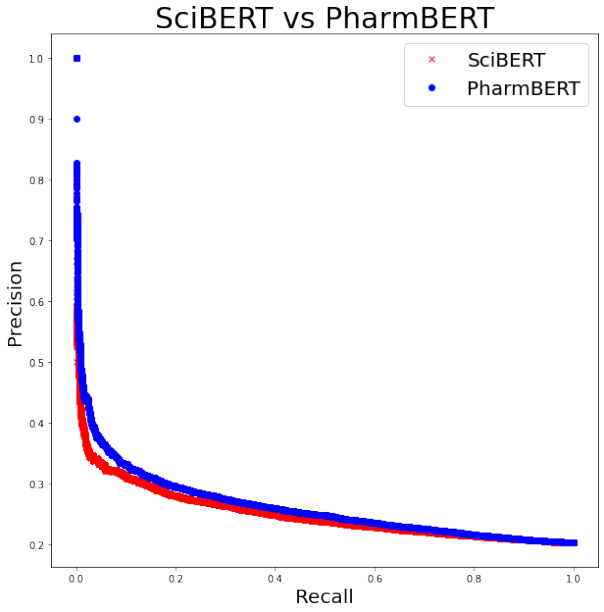

Figure 4: *DDI drug pair is predicted by calculating the distance between two embedded drug labels and see if it is less than a cutoff. When the cutoff is very small, the precision is high (all the predicted are correct) but the recall is low (only a small portion of DDI is predicted), when the cutoff is big, the recall is high (most of the DDI can be covered) but precision is low (a lot of misclassified example). From the curve, PharmBERT performed better than Scibert by producing higher precision results when the recall stays the same.*