W266 Final Project, Fall 2020
Valerie Meausoone, Prabhu Narsina

Leveraging Transformers and Transfer Learning for Relation eExtraction of Adverse Drug Events in Clinical Notes.

Abstract

The identification of Drugs that are tied to Adverse Drug Events (ADE) has implications for patient safety and clinical outcomes. With the rapid evolution of NLP, techniques using transformer models pre-trained on biomedical data can be adapted in a variety of ways to automate the detection of drug relations in clinical discharge notes, in particular ADE-Drug relations. We revisit a 2018 NLP challenge on ADE relation extraction with two deep learning approaches built on top of a pre-trained BioBERT model: pre-trained BioBERT fine-tuned with a fully connected layer and softmax activation (BioBERT+FC), and BioBERT with a Convolutional Neural Network (BioBERT+CNN). These two models managed to beat the 2018 baseline by 4 percentage points and reached an ADE-Drug relation F1 score of 0.87. We noted that the simpler architecture of BioBERT + FC required less fine-tuning to achieve high ADE-Drug relation F1 scores than BioBERT + CNN, confirming the power of using pre-trained Transformer architectures for downstream relation classification tasks. We note future directions for further improving F1 ADE-Drug relation scores. The source code is available here: https://github.com/pnarsina/ucb_mids_w266_fall2020_final/

Background

Clinical narratives contain detailed information on clinical events that is otherwise unavailable in structured Electronic Health Records. Natural Language Processing has enabled automating extracting information from clinical notes in a variety of ways. Named entity recognition and relation extraction stand as some of the most popular NLP tasks, allowing to study areas such as clinical decision making, mental health, and drug side-effect detection.[1] In particular, detecting Adverse Drug Events, defined as “injuries resulting from a medical intervention related to drugs and can include allergic reactions, drug interactions, overdoses, and medication errors”(https://health.gov/hcq/ade.asp), can have immediate impact in patient care and hospitalization outcomes.[2]To this end, the National NLP Clinical Challenges published in 2018 a new research data set and invited teams to work on an Adverse Drug Event and Medication extraction challenge consisting of a named entity recognition subtask (Task 1), and a relation extraction subtask (Task 2) [3]. For Task 2, this new research dataset consisted of clinical discharge summaries from MIMIC, annotated with eight relations: Strength-Drug, Form-Drug, Dosage-Drug, Frequency-Drug, Route-Drug, Duration-Drug, Reason-Drug, ADE-Drug.
NLP continues to evolve at a rapid rate and the models used by most task 2 teams in 2018 relied on extensive feature engineering (i.e.: distance between entities that make a relation, adding POS tags, etc.), in combination with NLP algorithms that are no longer considered gold standard today. Bi-Directional Long-Short Term Memory (LSTM), Convolutional Neural

Networks (CNN), among others, can today be replaced by or combined with the popular Bidirectional Encoder Representations from Transformers (BERT). In fact, pretrained BERT models have proven successful at relation extraction tasks, both on general benchmark datasets like TACRED [4,5], as well as on biomedical data. Architectures ranging from adding a simple fully-connected layer to adding a 1d-CNN model to pre-trained BioBERT, a BERT model pre-trained on biomedical corpora[8], have resulted in state-of-the-art F1 scores in the biomedical domain.[1] A handful of studies used pre-trained BERT models for extracting relations on the 2018 n2c2 data, reporting F-1 scores between 0.94 and 0.95[6,7]. To our knowledge, only one study focused on improving the performance of relation extraction specifically for Adverse Drug Events using BERT in combination with a particular sampling methodology (EDGE) and reported state-of-the-art performance, with an F1 score of 0.823[7].

We build on this previous work and further investigate how we can fine-tune pre-trained BioBERT to improve a relation extraction task focused on ADE-Drug performance. Our goal is to study how we can leverage a deep learning architecture to optimize ADE-Drug relation F1 scores, without the need for extensive data manipulation and feature engineering. As such, the contributions of our paper are as follows:

1. Revisiting the 2018 Relation-Extraction challenge on n2c2 data using successful implementations of pre-trained BERT models in the biomedical domain. Specifically, we test two implementations that have shown success in biomedical relation extraction: a pretrained BioBert model fine-tuned with a simply fully connected layer (BioBERT-FC) and a pretrained BioBert model fine-tuned with a Convolutional Neural Network (BioBERT-CNN).
2. Identifying how to fine tune BioBERT-FC and BioBERT-CNN to achieve the best performance on ADE-Drug relations.

For comparison with previous studies, F1 score serves as our primary evaluation metric. However, from a practical implementation standpoint, recall serves as our secondary evaluation metric, as it indicates whether our model is able to identify Adverse Drug Event relations.

Methods:

Data

The n2c2 dataset is available via the Data Portal Department of Biomedical Informatics (DBMI) in the Blavatnik Institute, at Harvard Medical School. It consists of 505 discharge summaries in text format, and 505 corresponding annotation files.

The dataset posed many challenges for appropriately handling the matching of relations to corresponding sentences in the discharge summaries. We did not find a transparent process for handling this in the literature and cover our approach in-depth in Appendix A. The average sequence length varied between 91 tokens for “no relation” and 255 tokens for Frequency-Drug, with long tails for all categories reaching thousands of tokens. The average sequence length for ADE-Drug sequences was of 210 tokens. We erred on the side of having longer sequences, given the inherently abbreviated and inconsistent nature of clinical discharge notes. Because many sections were structured as bullet points, we valued preserving related content in fewer sentences, rather than risking splitting relations that were laid out across multiple bullet points,

etc. With the exception of the Form-Drug relation, our distribution of drug relations tags, including the ADE-Drug relation, matched the distributions that were reported in the literature for n2c2 within ~5% [6]. Table 1 shows the breakdown of Drug-relation types in n2c2 data and highlights that the data is unbalanced. ADE-Drug relation class only represents 1% of all observations.

Relation Code	Number of observations
No Relation	68393 (53%)
Reason-Drug	8727 (7%)
Route-Drug	9648 (7%)
Strength-Drug	11055 (9%)
Frequency-Drug	10574 (8%)
Duration-Drug	1065 (1%)
Form-Drug	7667 (9%)
Dosage-Drug	7257 (5%)
ADE-Drug	1942 (1%)

Table 1: Drug-relation breakdown in n2c2 data

Our curated dataset had a total of 126,325 sentences, and 59,809 Drug relations. The n2c2 data was already pre-divided into a train and test datasets, which resulted in 76,315 and 50,010 sentences respectively. We further divided our train dataset into a train and validation datasets using an 80/20 split and resulting in 61,012 and 15,303 sentences each.

As described elsewhere[9], inputs for pre-trained BERT models fine-tuned on relation extraction tasks can be prepared in six different ways. After some experimentation on TACRED data, we assessed that including entity markers - SUB_S, SUB_E, OBJ_S, OBJ_E, generated the best F1 scores. We prepare our n2c2 data for BERT by including Subject start and end, and Object start and end markers. Subject markers are one of the Strength, Form, Dosage, Frequency, Route, Duration, Reason, ADE entities. Object markers are Drugs entities.

Deep-Learning Model

Our proposed approach focused less on feature engineering and more on testing a deep learning architecture that achieves state-of-the-art performance in extracting Drug relations, in particular Adverse Drug Event relations. As seen in Table 2, one of the challenges in identifying Adverse Drug Event-Drug relations is that ADE's are often reported in a subdued manner, and models fail to differentiate ADE-Drug relations and Reason-Drug relations [3,6,7]. The former is an injury resulting from a medical intervention related to a drug (https://health.gov/). The latter is a medical intervention involving a drug, intentionally used treat symptoms or a disease.

ADE-Drug Relation	Since no new infection was found this was presumed [**12-26**] steroids and the SUB_B leukocytosis SUB_E improved with OBJ_B prednisone OBJ_E taper.
Reason-Drug Relation	He also may have SUB_B recurrent seizures SUB_E which should be treated with OBJ_B ativan OBJ_E IV or IM and do not neccessarily indicate patient needs to return to hospital unless they continue for greater than 5 minutes or he has multiple SUB_B recurrent seizures SUB_E or complications such as aspiration.

Table 2: Examples of ADE-Drug and Reason-Drug relations.

Some subtle differences exist in the way ADE-Drug and Reason-Drug sentences are phrased. For example, we noted that the word “treat” appeared 943 times in Reason-Drug relation sentences, while it only appeared 89 times in ADE-Drug relation sentences.

As discussed earlier, the 2018 n2c2 challenge on relation extraction focused on ADE performance specifically, yet did not leverage what is considered today state-of-the-art BioBERT architecture. The best performing model from the 2018 n2c2 relations extraction challenge was an ensemble model composed of a CRF, BiLSTM-CRF, and ADDRESS, a BiLSTM-CRF–based joint topic-relation extraction method, put out by UTHealth/Dalian (UTH). We refer to this model as the UTH model and consider this our baseline, since they achieve an overall F1 score of 0.9630, and an F1 score on ADE-Drug relation of 0.822.

In this study, we present two methods for fine-tuning BioBERT models that have shown good performance on relation extraction tasks, with the goal of recognizing differences in intent between ADE-Drug relations and Reason-Drug relations. As shown in Figure 1a, the first method consists of adding a fully-connected layer and softmax function to a pre-trained BioBERT. Building on the principle that “Attention is all you need”[11], we hypothesize that BioBERT’s Transformer’s architecture can adequately capture the context in each n2c2 sequence and can successfully maximize the log-probabilities generated by BioBERT’s CLS tokens (“C” - Class label) in the final layer. The second method, also shown on Figure 1b, is based on a more complex architecture, which utilizes all of BioBERT’s final hidden state vectors and implements a 1-d Convolutional Neural Network to further extract hyper-features, before classifying each n2c2 sequence using a fully connected layer and softmax function. This architecture has been previously reported in the literature for relation extraction in biomedical data [1,12,] and we hypothesize that Convolutions may highlight some local text patterns that further differentiate ADE-Drug from other classes.

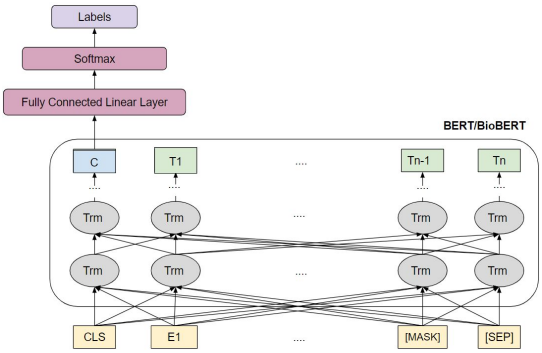

Figure 1a: BioBERT + FC

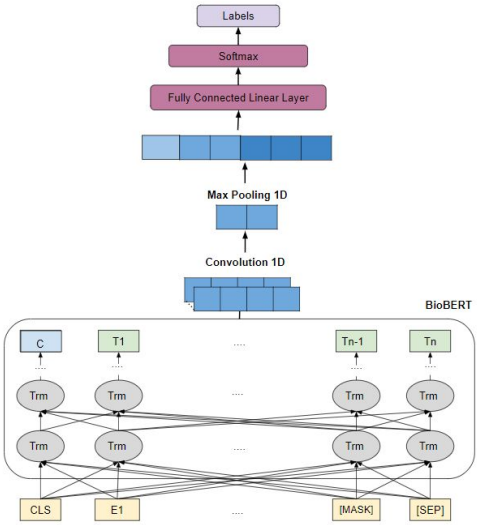

Figure 1b: BioBERT + CNN

For each proposed model (Figures 1a and 1b), we anticipate the need to include class weights that emphasize the loss of the ADE-Drug class in our Cross Entropy loss function, given that our dataset is imbalanced. We also assess how hyperparameter tuning improves the F1 score of the ADE-Drug relation. While not disclosed in this paper, we ran over 50 experiments allowing us to assess the impact of hyper parameters such as number of frozen layers, sequence length, training batch size, learning rate, number of warm up steps, softmax weights, activation, and kernel number and sizes (for CNN). We also ran iterations with binarized class outcomes and attempted to develop a custom softmax implementation that favored ADE-Drug relations, which were both inconclusive in that they produced low overall F1 scores. In the next section, we present our best model implementations.

Results

After hyperparameter tuning, BioBERT+FC and BioBERT-CNN produced comparable results in terms of overall F1 score, both slightly exceeding the UTH baseline. Similarly, both methods also achieved an F1 score of 0.87 for ADE-Drug relation, beating the baseline of 0.82. We included an additional iteration, which showed a high recall score of 0.91 and we argue has relevance in the clinical setting.

Table 3 below presents the hyperparameters that provided optimal results in both BioBERT+FC and BioBERT+CNN configurations.

Best Model	Hyperparameters	Overall F1 Score	ADE-Drug F1 Score	ADE-Drug Recall
UTH (Baseline)	Not applicable	0.9630	0.82	Not disclosed

BioBERT + FC	Batch size = 12 Sequence length= 384 ADE weight= none Learning Rate= 1e-5 Dropout = 0.1 Optimizer= Adam	**0.9639**	**0.87**	0.84
BioBERT + CNN	Sequence length= 256 Kernel number= 1 Kernel size= 3 Stride= 1 ADE weight= 8 Learning Rate= 1.25e-5 Layer Normalization Optimizer= Adam	0.9546	0.81	**0.91**
BioBERT + CNN	Sequence length= 256 Kernel number= 1 Kernel size= 24 Stride= 1 ADE weight= 8 Learning Rate= 1.25e-5 Layer Normalization Optimizer= Adam	**0.9620**	**0.87**	0.87

Table 3: Best performing models for overall Drug relation classification, and ADE-Drug classification on n2c2 data.

While BioERT+FC and BioBERT+CNN both managed to beat the baseline, BioBERT+FC proved to be more consistent in producing F1-scores for ADE-Drug that were above 0.80 for a variety of hyperparameter combinations.

Effect of class weights:

Contrary to our assumptions, including class weights in our Cross Entropy loss function did not make a difference in ADE-Drug F1 scores for BioBERT+FC. When using a weight of 4, the recall for the ADE-Drug relation improved recall from 0.84 to 0.87 and at the expense of precision,which decreased from 0.91 to 0.89. When increasing the weights past 4, precision, recall and f1-scores for the ADE-Drug relation increased on the train data and decreased on the test data, indicating overfitting. In contrast, BioBERT+CNN was not successful in our ADE-Drug relation extraction task unless we included class weights to emphasize the ADE in our Cross Entropy loss function. Not including any weights led to overall scores of 0 for precision, recall and F1. A weight of 8 was required to achieve ADE-Drug F1 scores similar to the BioBERT + FC model. Figure 2 below shows how including weights in BioBERT + CNN helped stabilize and optimize the learning of our model. Because the number of parameters is greater in the BioBERT + CNN model than in BioBERT+FC, we believe that classes having insufficient data must be weighted in BioBERT + CNN in order to produce stable results.

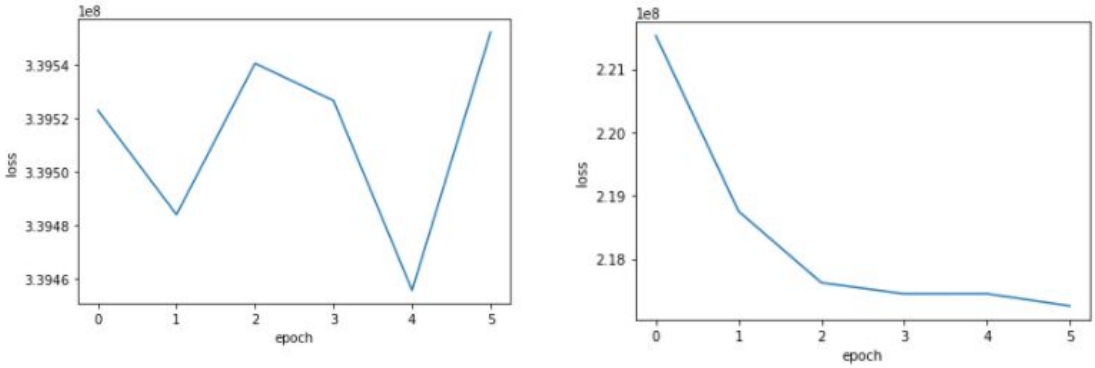

Figure 2: Cross Entropy Loss versus number of epochs when using no weights (left) and when using a weight of 8 (right) for ADE-Drug relation in BioBERT+CNN.

Effect of kernel size and number:

For our BioBERT+CNN model, best F1 and recall scores for the ADE-Drug relation were obtained when using only one kernel. We tested including up to three parallel 3 kernels and max pooling combinations, as well as generating deeper CNN models with 3 consecutive kernel and max pooling combinations, and these attempts resulted in producing F1 scores for ADE-Drug relation between 0.82 and 0.84 without substantially improving precision or recall. Furthermore, we noticed that deeper CNN networks weighted on ADE-Drug class were not successful on other classes with a low number of observations, bringing to 0 the F1 score of the Duration-Drug relation. Using a small kernel of size 3 provided some of the highest recall results, as seen in table 3. Using a larger kernel of size 24 improved precision at the expense of recall, and as a result improved the ADE-Drug relation F1 score.

Effect of sequence length:

Our original decision of generating longer sentences when tokenizing discharge summaries led us to believe that including larger sequence lengths would positively impact the performance of our model. This was partially true, as for both models the impact of changing a sequence length from 128 to 256 was most noticed in measures other than the model F1-scores. For BioBERT+FC, recall tended to improve with increased sequence length, while for BioBERT+CNN, it was precision. We noted that increasing sequence length to 384 improved the overall F1 score of the BioBERT+FC but not the ADE-Drug relation F1 score. We speculate that it would have had the same impact on BioBERT+CNN, but did not run this iteration due to a lack of time and computational resources. For the purpose of improving ADE-Drug relation performance, since the average ADE-Drug sequence length was of 210 tokens, we argue that using sequence length of 256 is appropriate and more computationally efficient.

Effect of freezing BERT pre-trained layers:

As reported in the literature, freezing the bottom layers of our BERT pre-trained model makes most intuitive sense as they are thought to attend broadly while the top layers capture linguistic

syntax [10]. Our iterations indicated that this provided optimal results for BioBERT+CNN, but that actually freezing all the layers worked best on BioBERT+FC.

Misclassified ADE-Drug relations:
Across our experiments, we looked at the top 100 ADE-Drug relations that got frequently misclassified by BioBERT + FC and BioBERT + CNN. These sequences were on average shorter (144 tokens compared to the overall ADE-Drug relation length of 209 tokens). Upon inspection, we noticed that these sequences were either ambiguous because they did not exclusively address patient symptoms due to an ADE event (i.e.: " Patient relatively immunocompromised due both to cancer diagnosis and chronic OBJ _ B steroid OBJ _ E use secondary to SUB _ B brain tumor SUB _ E") . Other sentences showed abbreviations or were so short that we hypothesized that context could not easily be picked up, even by a pre-trained BioBERT model (i.e.: "OBJ _ B Ceftriaxone OBJ _ E / Rifaximin was given immunosuppression and SUB _ B elevated WBC SUB _ E .").

Discussion

Our study confirmed the power of pretrained domain-specific BERT models applied to downstream tasks such as relation classification. While we had originally anticipated having to carefully fine-tune our proposed models to optimize ADE-Relalation, we found overall that simpler architectures on top of pretrained BERT perform just as well, if not better, than more complex ones. While we managed to find implementations of both models that beat the baseline reported in 2018 for the ADE-Drug relation F1 score, BioBERT+FC was much easier to implement than BioBERT+CNN, and appeared more robust to hyperparameter tuning. For BioBERT+CNN, the implementation of one simple kernel was the model that produced better results, compared to implementations with more kernels, and with deeper networks.
We were surprised by the effect, or lack thereof, of adding class weights to our Cross Entropy Loss function on the ADE-Drug relation F1 score for BioBERT + FC. It appeared that including weights had a stronger impact on related measures, such as precision and recall, but that overall F1 scores ended up evening out. In contrast, we noted the huge effect of class weights on the BioBERT+CNN model. We suggest that BioBERT+CNN may be more sensitive to class imbalance than the BiobERT+FC, because the former model introduces new parameters that need to be trained and need to rely on sufficient data for each n2c2 class size.
Because our BioBERT+FC model is a slightly fine-tuned version of the pretrained BioBERT model, weighting and other types of hyperparameter tuning may not have as much impact on maximizing the log probabilities of the [CLS] token ("C" - Class label) in the final layer.
We nevertheless found an implementation of BioBERT+CNN that produced higher recall scores for ADE-Drug relation. While it is possible that a similar outcome could have been achieved using a yet unexplored combination of hyperparameters on BioBERT+FC, our BioBERT+CNN is a simple enough architecture to be considered relevant in the clinical domain. The focus of this study has mostly been on achieving the best F1 score, as it is a sensible metric that balances precision and recall and is the most common metric reported for ADE-Drug relation in the literature. However, from a patient safety perspective, one could argue that favoring recall could help with understanding the gamut of drug events that can be harmful to patients.

The focus of this study was to maximize the F1 score for the ADE-Drug relation, yet we found that our numerous experiments consistently showed a lower ADE-Drug relation score compared to other classes. Based on our misclassification analysis, a future direction on this work could continue building on pre-trained BioBERT models and incorporate tags in the input n2c2 sequences that emphasize the ADE-Drug relation. Because of the inconsistent quality of text in discharge notes (e.g.: use of abbreviations, incomplete sentences, etc.) adding features in combination with a pretrained BioBERT model could enhance the F1 score of ADE-Drug relations. Finally, a direction could be to revisit our failed attempts at generating a custom softmax activation, such as large margin softmax, that could increase the distance between classes and minimize the distance between the same classes.[13]

Conclusion

This study focused on revisiting a 2018 NLP challenge, which sought to maximize F1 scores on drug relation extraction in clinical notes. We tested two architectures on top of a pre-trained BioBERT model, with one one that consisted of a simple fully connected layer and softmax (BioBERT + FC). The other leveraged the full dimensionality of the last hidden layer of BioBERT by including a 1-d Convolutional Neural Network, with one kernel (BioBERT+CNN). Our two implementations managed to match the overall F1 score of the winning team in 2018, and beat their ADE-Drug relation F1 scores. We note however that the simple architecture of BioBERT + FC is more resistant to changes in hyperparameter tuning and achieved comparable results compared to the more complex BioBERT+CNN.

References

1. Tao Chen, Mingfen Wu, Hexi Li, A general approach for improving deep learning-based medical relation extraction using a pre-trained model and fine-tuning, Database, Volume 2019, 2019, baz116, https://doi.org/10.1093/database/baz116
2. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics. [Internet]. 2009 Oct [cited 2019 Mar 11];42(5):760-72.
3. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020 Jan 1;27(1):3-12. doi: 10.1093/jamia/ocz166
4. Yamada, Ikuya & Asai, Akari & Shindo, Hiroyuki & Takeda, Hideaki & Matsumoto, Yuji. (2020). LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention.
5. Yang, SungMin & Yoo, SoYeop & Jeong, OkRan. (2020). DeNERT-KG: Named Entity and Relation Extraction Model Using DQN, Knowledge Graph, and BERT. Applied Sciences. 10. 6429. 10.3390/app10186429.
6. Wei Q, Ji Z, Si Y, Du J, Wang J, Tiryaki F, Wu S, Tao C, Roberts K, Xu H. Relation Extraction from Clinical Narratives Using Pre-trained Language Models. AMIA Annu Symp Proc. 2020 Mar 4;2019:1236-1245.

7. Guan H, Devarakonda M. Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes. AMIA Annu Symp Proc. 2020;2019:1051-1060. Published 2020 Mar 4.
8. BioBERT: 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*, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682
9. Livio Baldini Soares, Nicholas Arthur FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In Association for Computational Linguistics (ACL), pages 2895–2905.
10. Lee, Jaejun & Tang, Raphael & Lin, Jimmy. (2019). What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning.
11. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.
12. Zheng, Shaomin & Yang, Meng. (2019). A New Method of Improving BERT for Text Classification. 10.1007/978-3-030-36204-1_37.
13. Liu, Weiyang & Wen, Yandong & Yu, Zhiding & Yang, Meng. (2016). Large-Margin Softmax Loss for Convolutional Neural Networks. ProC. Int. Conf. Mach. Learn..

Appendix A

Each annotation file contains tagged entities related to drug name, frequency, dosage, strength, route, duration, reason and ADE's, as well as tagged relations between these entities.
Within an annotation file, each entity is defined on a separate row by its raw name as well as its start and end position in the discharge summary text file. Each relation is defined by a pair of tagged entities and one of the relation types defined above. Figure 1 below shows an excerpt of the annotation file.

```
T1    Reason 10179 10197 recurrent seizures
R1    Reason-Drug Arg1:T1 Arg2:T3
T3    Drug 10227 10233  ativan
```

Figure 1: Annotation file sample from the n2c2 dataset

Our goal was to prepare data for BERT. We lay out our process below.

Sentence tokenization
Discharge summaries were tokenized into sentences using NLTK's sentence tokenizer function. While in general, these summaries were organized in similar sections, each discharge summary could be structured in slightly different ways. Furthermore, unlike regular written English, discharge summaries contained a lot of bullet point phrases that did not start with a capital letter and ended with a dot. Upon inspection, NLTK's sentence tokenizer helped identify main traditional sentences, but large chunks of text remained combined into one tokenized sentence. Based on n2c2's documentation, relations were tagged using the logic of the most proximal location in text. That is, if a drug relation was identified within the same sentence, that is the

relation tag that was included in the annotation file. If no same-sentence relation was found, annotators proceeded by looking at the two closest sentences, and so forth.
For this reason, we decided to further process our tokenized sentences by splitting paragraphs separated by two line breaks or more. However, discharge summary phrases were sometimes artificially separated by single line breaks or bullet points, so we decided *not* to split sentences further in order to preserve semantic cohesion.

Matching of sentences and relations.
For each discharge summary, we then processed the annotation files into a dictionary containing as keys the two tags making up a given drug relation and as values the specific relation. Using the example from Figure 1 above, an entry in the dictionary for a specific discharge summary would be { (“recurrent seizures”, “ativan”) : “Reason-Drug” }. For each discharge summary and its corresponding dictionary, we assessed whether an element of a key (e.g.: “recurrent seizures”) was contained within a tokenized sentence. If an element was found in a given sentence, we temporarily retained its sentence’s index.We repeated this process for the other element in the key (e.g.: “ativan”). Based on n2c2’s annotation documentation, we then selected index pairs with the minimum difference. If two indices ended up in the same sentence, the sentence was matched to the relation. If two indices ended up in two distinct sentences, we concatenated the sentences. Sequences that were not matched with a relation were labeled with the relation “no relation”.

Adding positional tags
Based on the literature[9], we added positional markers - subject’s start and end, as well as an object’s start and end - corresponding to the two tagged entities making up a relation.

We saved each original or concatenated sentence with its corresponding drug relation, and processed the data further to make it compatible for BERT: we generated an index column, an alpha column, an encoded relation code column and a last column with the actual string sequence.

Relation Code	**Number of observations**	
	Method reported by Wei et al. (2020) N	**Our method N(% difference)**
Reason-Drug	8578	8727 (-2%)
Route-Drug	9084	9648(-6%)
Strength-Drug	10946	11055 (-1%)
Frequency-Drug	10344	10574 (-2%)
Duration-Drug	1069	1065 (0%)

Form-Drug	11028	7667 (44%)
Dosage-Drug	6920	7257 (-5%)
ADE-Drug	1840	1942 (-5%)