Text Mining Drug/Chemical-Protein Interactions using an Ensemble of BERT and T5 Based Models

Virginia Adams[§], Hoo-Chang Shin[§], Carol Anderson[§], Bo Liu[§], Anas Abidin[§],
NVIDIA / Santa Clara, California, USA
{vadams;hshin;carola;boli;aabidin}@nvidia.com

***Abstract*—In Track-1 of the BioCreative VII Challenge participants are asked to identify interactions between drugs/chemicals and proteins. In-context named entity annotations for each drug/chemical and protein are provided and one of fourteen different interactions must be automatically predicted. For this relation extraction task, we attempt both a BERT-based sentence classification approach, and a more novel text-to-text approach using a T5 model. We find that larger BERT-based models perform better in general, with our BioMegatron-based model achieving the highest scores across all metrics, achieving 0.74 F1 score. Though our novel T5 text-to-text method did not perform as well as most of our BERT-based models, it outperformed those trained on similar data, showing promising results, achieving 0.65 F1 score. We believe a text-to-text approach to relation extraction has some competitive advantages and there is a lot of room for research advancement.**

Keywords*— *BERT, BioBERT, BioMegatron, T5, Text-to-Text

I. Introduction

The task of relation extraction, particularly for drug/chemical and protein, can be useful for many applications. For instance, finding if a chemical is a down-regulator or up-regulator of a protein can be useful for drug discovery. Christopoulou et al. (*1*) show using relation extraction for finding adverse drug events.

With this motivation, the DrugProt shared task in BioCreative VII challenge/workshop is to find the relation between drug/chemical and proteins from biomedical literature in PubMed. The task is to classify drug/chemical and protein relation into 13 possible classes, where the named entities of drug/chemical and proteins (candidates for the relations) are provided as annotated.

There was a previous related ChemProt task (*2*), where the data format and task are similar. DrugProt task has more data and the relation annotations are more granular. Since the introduction of the ChemProt task, Lee et al. (*3*) and many others have shown the effectiveness of BERT (*4*) models pre-trained in-domain PubMed data. In addition, Beltagy et al. (*5*) and Gu et al. (*6*) show the benefits of in-domain vocabulary set learned from PubMed literatures.

Furthermore, Shin et al. (*7*) show the additional benefit of larger model size with their BioMegatron models. Here, the authors show that larger domain-specific language models outperform their smaller out-of-domain counterparts on a variety of biomedical natural language processing tasks. For our BioCreative VII Track-1 submission, we first repeat the BioMegatron study with the given Track-1 data, verifying the size and domain-specific hypothesis.

We then experiment with a novel text-to-text approach using T5 (*8*). We convert the relation extraction task from sentence classification into a text-based question answering problem. We introduce novel ideas such as multi-step question answering and question balancing to improve performance.

Lastly, we use a model ensemble technique (*9*) to boost the final performance of our submissions.

[§]In reverse alphabetical order - authors contributed equally.

II. Dataset, Pre-Processing and Software

The DrugProt dataset (*10*) contains abstract, named entities of drug/chemical and protein pairs. In the training and development set, the relation annotation for the pairs are also provided. We conduct minimal pre-processing where we *(i)* break the abstracts into sentence-level, and *(ii)* sub-tokenize the words as in (*4*, *8*).

We use Pandas data library [1] for pre-processing, and PyTorch-based Megatron-LM (*11*) and NeMo [2] codebase for further pre-processing, training, and testing.

III. BERT-Based Models

Our BERT-based models use the widely adopted relation extraction approach of annotating entities with special tokens and performing sentence classification. Formulated as multi-class classification, there are 14 total classes – 13 relation classes and an additional no-relation class. We also try the recent advanced method of "matching the blanks" (*12*) using open-sourced code (*13*), but find the benefits of model size and a domain-specific pre-training corpus outweigh the benefit of a novel training scheme.

Examples of converting the original text and dataset for multi-relation sentence classification is shown in Figure 1. We break each abstract into sentences and annotate entity pairs of interest with special tokens to facilitate relation extraction as a sentence classification task. We experiment with two slightly different entity annotation schemes: *(i)* masking the entities with special chemical and gene tokens (*14*), and *(ii)* surrounding the entities with chemical and gene tags (*15*).

We attach a fully-connected layer to the final BERT pooling layer for the classification. We experiment with {1, 2} number-of-layers and {128, 256, 384, 512} number of hidden-units for the fully-connected layer. Other hyper-parameters we experiment are: max-sequence-length of {128, 256, 512}, dropout-rate of {0.1, 0.5, 0.9}, learning-rate of {5e-6, 9e-6, 1e-5, 5e-5}. We use Adam optimizer (*16*) with cross-entropy loss and train for 5 to 40 epochs. The BERT-cased vocabulary (*4*) is used due to the time limitation to conduct enough experiments by the submission deadline.

Development set BERT-based evaluation results are shown in Table I. Because the methods, vocabulary sets, and hyper-parameters are not consistent, it is not a completely controlled experiment. Nonetheless, we can observe the general trend of *(i)* domain-specific models (BioBERT, BioMegatron) and *(ii)* larger model size contributing beneficially to improved performance.

IV. T5-Based Models

While BERT-based models have been a standard for some time now, more recent "text-to-text" or "prompting" based methods (*8*, *17*) have several advantages over transformer encoder language models with additional task specific architectures. One of the most notable advantages is the ability to perform multiple tasks without needing distinct specialized layers for each task. Through this formulation,

[1]https://pandas.pydata.org/

[2]https://github.com/NVIDIA/NeMo

Abstract

RDH12, a retinol dehydrogenase causing Leber's congenital amaurosis, is also involved in steroid metabolism. Three retinol dehydrogenases (RDHs) were tested for steroid converting abilities: human and murine RDH 12 and human RDH13. RDH12 is involved in retinal degeneration in Leber's congenital amaurosis (LCA). We show that murine Rdh12 and human RDH13 do not reveal activity towards the checked steroids, but that human type 12 RDH reduces dihydrotestosterone to androstanediol, and is thus also involved in steroid metabolism. Furthermore, we analyzed both expression and subcellular localization of these enzymes.

TreeLSTM format

17512723 We show that BC6OTHER and BC6ENTG do not reveal activity towards the checked steroids, but that BC6OTHER reduces dihydrotestosterone to BC6ENTC , and is thus also involved in steroid metabolism. False CPR:0 T1 androstanediol T7 human RDH13

17512723 We show that BC6OTHER and BC6OTHER do not reveal activity towards the checked steroids, but that BC6ENTG reduces dihydrotestosterone to BC6ENTC , and is thus also involved in steroid metabolism. True CPR:12 T1 androstanediol T9 human type 12 RDH

SemEval2010 format

17512723 We show that murine Rdh12 and <eg>human RDH13</eg> do not reveal activity towards the checked steroids, but that human type 12 RDH reduces dihydrotestosterone to <ec>androstanediol</ec> , and is thus also involved in steroid metabolism. False CPR:0 T1 androstanediol T7 human RDH13

17512723 We show that murine Rdh12 and human RDH13 do not reveal activity towards the checked steroids, but that <eg>human type 12 RDH</eg> reduces dihydrotestosterone to <ec>androstanediol</ec> , and is thus also involved in steroid metabolism. True CPR:12 T1 androstanediol T9 human type 12 RDH

Fig. 1. An example converting the original text and dataset for multi-relation sentence classification into our BERT fine-tuning format. In the TreeLSTM (*14*) format, we mask the entities of interest with special chemical (BC6ENTC) and gene (BC6ENTG) tokens, while masking other chemicals and genes not under consideration with the special "other" token (BC6OTHER). False CPR:0 indicates there is no relationship between the two entities under examination while True CPR:12 signifies the true relationship's index is twelve, corresponding to "product-of". In these examples T1 indicates androstanediol, the first entity to appear in the abstract, and T9 human type 12 RDH is the ninth entity in the abstract. In the SemEval2010 (*15*) format, we surround the entities of interest with special chemical (<ec>human type 12 RDH</ec>) and gene (<eg>androstanediol</eg>) tags, leaving all other chemicals and genes found in the sentence unannotated.

Text-to-text format

is there a relationship between "androstanediol" and "human type 12 RDH" from context "We show that murine Rdh12 and human RDH13 do not reveal activity towards the checked steroids, but that human type 12 RDH reduces dihydrotestosterone to androstanediol, and is thus also involved in steroid metabolism."? answer: "yes"

predict the relationship between "androstanediol" and "human type 12 RDH" in context "We show that murine Rdh12 and human RDH13 do not reveal activity towards the checked steroids, but that human type 12 RDH reduces dihydrotestosterone to androstanediol, and is thus also involved in steroid metabolism.". answer: "product-of"

Fig. 2. An example of converting the original text and dataset into the text-to-text T5-base fine-tuning and evaluation format. The abstract is split into sentences. Each sentence is transformed into a sequence of question and answer pairs: *first*, asking if a relationship between two specific entities is present in the sentence, and *second*, prompting the model to predict the relationship if it indicated one exists.

every task can share the same cross-entropy minimization training objective and the inductive bias learned about biomedical entities for one task can possibly be transferred to another.

A further advantage is the potential for "zero-shot" or "few-shot" learning (*17*). When tasks are expressed as text-based question answering problems, with sufficient model capacity, text-to-text generative models can produce impressive results on multiple tasks with little or no task-specific fine-tuning (*18*). Though we fine-tune on the challenge data training set for this submission, we view this as the necessary first step in working towards our end goal of few-shot or zero-shot relation extraction.

Many natural language processing tasks have been successfully reformulated as text-to-text tasks, such as text classification, natural language inference, summarization, and reading comprehension. To our knowledge there are no published studies to date that use a text-to-text approach for relation extraction, although a prompting-based approach using masked language modeling has been explored by Wei et al. (*19*).

We use T5 (*8*) for our text-to-text approach. Figure 2 shows an example of data conversion from the given biomedical abstract and entity annotations into the T5 prompting input and output. The abstract is split into sentences, and each sentence is turned into a sequence of natural language questions and answers. We first ask the model to identify if a specific relation is present in the sentence. If there is, we ask it to predict the relation. We investigated different prompt formats and empirically found this setup to yield the highest scores. We fine-tuned an off-the-shelf T5-base model that was pre-trained on general domain text via the methodology described by Raffel et al. (*8*).

Evaluation results using T5 on the development set are shown in Table I. A noticeable improvement is achieved by balancing the positive and negative examples of sentences with and without relations and then over sampling the number of negative examples in the training set. Our best T5 model out performed our BERT-Base with BERT-uncased vocabulary model and performed within 0.01 F1 score of our BERT-Base with cased vocabulary model. BERT-Large models and BERT-Base models pre-trained on biomedical domain data out-score fine-tuned T5, but perhaps with in-domain pre-training and larger model capacity, T5 could further improve.

Recent studies (*18*, *20*) show that model size needs to be sufficiently large, such as having 5 billion parameters, in order to achieve good zero-/few-shot performance. Since our T5 models were relatively small (345 million parameters), we will definitely need to increase model size for few or zero shot relation extraction to be

TABLE I
PRECISION, RECALL, AND F-1 SCORES OF T5 AND BERT MODELS ON THE DEVELOPMENT SET.

Model	Method	#Parameters	Vocabulary	Prec	Rec	F-1
BERT-base	(*12*)	110m	BERT-uncased	0.68	0.59	0.63
BioBERT-base	(*12*)	110m	BERT-uncased	0.71	0.67	0.69
BERT-base	—	110m	BERT-cased	0.77	0.58	0.66
BERT-large	—	345m	BERT-cased	0.74	0.70	0.72
BioMegatron	—	**345m**	**BERT-cased**	**0.76**	**0.71**	**0.74**
T5-base	over sampling positive	345m	BERT-uncased	0.36	0.78	0.49
T5-base	balancing negative/positive	345m	BERT-uncased	0.54	0.71	0.61
T5-base	**over sampling negative**	**345m**	**BERT-uncased**	**0.67**	**0.63**	**0.65**

feasible. Nevertheless, our results are encouraging.

V. MODEL ENSEMBLE AND FINAL TEST-SET SCORES

Model ensembling (*9*) is a widely used technique to increase the final performance of machine learning models by combining multiple models' predictions, often averaging them.

For our final submission, we use an ensemble of different models and attain noticeable improvement in evaluation scores on both the development and test sets. We ensemble the models by taking a weighted average of each model's predicted probability vector then selecting the argmax from this averaged vector as our final prediction. This approach can work across a diverse set of models, even between our BERT and text-to-text models. In fact, ensembling diverse models is desirable because each single model's unique prediction errors can be overcome by generally low probability scores from the other models in the ensemble, masking individual model mistakes.

We make four submissions in total. Ordered as in Table II, our first submission is an ensemble of our fine-tuned BioBERT-Base and best T5 models. Our second submission is from our best T5 model alone. Third, we submit single model predictions from fine-tuned BioMegatron. Finally, our fourth and best submission as an ensemble of fine-tuned BERT-Base, BERT-Large, and BioMegatron.

TABLE II
FINAL PRECISION, RECALL, AND F-1 SCORES ON THE TEST SET.

Model(s)	Prec	Rec	F-1
[BioBERT, T5]-ensemble	0.71	0.67	0.69
T5	0.64	0.58	0.61
BioMegatron	0.74	0.72	0.73
[BERT-base&large, BioMegatron]-ensemble	**0.77**	**0.73**	**0.75**

Table II shows the final official evaluation scores on the test set. The best scores are achieved using an ensemble of multiple BERT-base&-large, and BioMegatron models. Table III shows the final official evaluation scores on the test set at a more granular level for each relation type. Some relations with only a few samples in the training data are generally difficult to classify correctly.

VI. LARGE-SCALE SUB-TRACK

For the large-scale track we did not perform additional model development due to time constraints. We only remove pipelines unnecessary for inference from our smallest BERT-base model. This mostly includes convenience pipelines in PyTorch-Lightning. We then run inference on four GPUs, dividing the dataset into four sub-parts. Our resulting BERT-base model reports lower evaluation scores on the development set. It is possible we lost some performance due to lack of attention in model-stripping. Nonetheless, we finished inference on the large-scale sub-track test data. The overall- and granular- official evaluation scores for the large-scale sub-track are shown in Table IV and Table V.

TABLE III
GRANULAR SCORES FOR EACH RELATION TYPE FOR TWO OF OUR BEST-PERFORMING ENSEMBLE MODELS.

	[BioBERT, T5]			[BERT, BioMegatron]		
Relation-Type	**Prec**	**Rec**	**F-1**	**Prec**	**Rec**	**F-1**
ACTIVATOR	0.75	0.64	0.69	0.77	0.76	0.77
AGONIST	0.81	0.73	0.77	0.76	0.722	0.74
AGONIST-INHIBITOR	1.0	0.33	0.5	0.0	0.0	0.0
ANTAGONIST	0.84	0.87	0.85	0.85	0.92	0.88
DIRECT-REGULATOR	0.68	0.56	0.62	0.73	0.65	0.69
INDIRECT-DOWNREGULATOR	0.62	0.77	0.69	0.74	0.76	0.75
INDIRECT-UPREGULATOR	0.71	0.67	0.69	0.76	0.78	0.77
INHIBITOR	0.75	0.77	0.76	0.84	0.84	0.84
PART-OF	0.64	0.66	0.65	0.72	0.59	0.65
PRODUCT-OF	0.64	0.59	0.61	0.70	0.56	0.62
SUBSTRATE	0.63	0.46	0.53	0.68	0.53	0.59
SUBSTRATE_PRODUCT-OF	0.0	0.0	0.0	0.0	0.0	0.0
AGONIST-ACTIVATOR	0.0	0.0	0.0	0.0	0.0	0.0

TABLE IV
FINAL PRECISION, RECALL, AND F-1 SCORES ON THE LARGE-SCALE SUB-TRACK.

Model(s)	Prec	Rec	F-1
BERT-base	0.73	0.33	0.46

VII. DISCUSSION

Our results confirm previous findings that larger models tend to perform better than smaller ones, and models trained on domain-specific text tend to perform better than those trained on general domain data. For our BERT-based models, performance could potentially be improved beyond what we reported here by using even

TABLE V
GRANULAR SCORES FOR EACH RELATION TYPE ON THE LARGE-SCALE SUB-TRACK.

Relation-Type	Prec	Rec	F-1
ACTIVATOR	0.69	0.35	0.46
AGONIST	0.74	0.40	0.52
AGONIST-INHIBITOR	0.0	0.0	0.0
ANTAGONIST	0.75	0.48	0.58
DIRECT-REGULATOR	0.66	0.19	0.30
INDIRECT-DOWNREGULATOR	0.71	0.35	0.47
INDIRECT-UPREGULATOR	0.70	0.43	0.53
INHIBITOR	0.81	0.45	0.58
PART-OF	0.60	0.15	0.25
PRODUCT-OF	0.62	0.23	0.33
SUBSTRATE	0.69	0.17	0.27
SUBSTRATE_PRODUCT-OF	0.0	0.0	0.0
AGONIST-ACTIVATOR	0.0	0.0	0.0

larger BioMegatron models, which we did not have time to complete and submit. For our T5 models, larger model size and pretraining on in-domain text would likely improve performance. We also confirm that model ensembling gives an additional performance boost, even when model architectures are different (BERT- and T5- based).

Although our text-to-text based methods did not perform as well as the largest BERT models we trained, their performance was similar to or better than BERT base models pretrained on general domain text. These results indicate that relation extraction can be successfully framed as a text-to-text task, while also highlighting some challenging aspects of the approach. In particular, we find that careful attention should be paid to class balancing during fine-tuning and to the design of prompts used for inference.

For the large-scale sub-track, we could only use our smallest BERT-base model. Further improvement could be seen by applying advanced model optimization techniques such as quantization and pruning, allowing use of our larger models for the large-scale inference task.

REFERENCES

1. F. Christopoulou, T. T. Tran, S. K. Sahu, M. Miwa, S. Ananiadou, *Journal of the American Medical Informatics Association* **27**, 39–46 (2020).
2. O. Taboureau *et al.*, *Nucleic acids research* **39**, D367–D372 (2010).
3. J. Lee *et al.*, *Bioinformatics* **36**, 1234–1240 (2020).
4. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *arXiv preprint arXiv:1810.04805* (2018).
5. I. Beltagy, K. Lo, A. Cohan, *arXiv preprint arXiv:1903.10676* (2019).
6. Y. Gu *et al.*, *ACM Transactions on Computing for Healthcare (HEALTH)* **3**, 1–23 (2021).
7. H.-C. Shin *et al.*, *arXiv preprint arXiv:2010.06060* (2020).
8. C. Raffel *et al.*, *arXiv preprint arXiv:1910.10683* (2019).
9. G. Hinton, O. Vinyals, J. Dean, *arXiv preprint arXiv:1503.02531* (2015).
10. A. Miranda *et al.*, *Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations*, 2021.
11. M. Shoeybi *et al.*, *arXiv preprint arXiv:1909.08053* (2019).
12. L. B. Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, *arXiv preprint arXiv:1906.03158* (2019).
13. https://github.com/plkmo/BERT-Relation-Extraction, [Online; accessed 1-Sep-2021].
14. S. Lim, J. Kang, *Database* **2018** (2018).
15. I. Hendrickx *et al.*, *arXiv preprint arXiv:1911.10422* (2019).
16. D. P. Kingma, J. Ba, *arXiv preprint arXiv:1412.6980* (2014).
17. P. Liu *et al.*, *arXiv preprint arXiv:2107.13586* (2021).
18. T. B. Brown *et al.*, *arXiv preprint arXiv:2005.14165* (2020).
19. X. Chen *et al.*, *KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction*, 2021, arXiv: 2104.07650 (cs.CL).
20. J. Wei *et al.*, *arXiv preprint arXiv:2109.01652* (2021).