A Multi-Task Transfer Learning-based method for extracting Drug-Protein Interactions

Ed-drissiya El-allaly[1], Mourad Sarrouti[3], Noureddine En-Nahnahi[1], Said Ouatik El Alaoui[2]
[1]Laboratory of Informatics, Signals, Automatic, and Cognitivism (LISAC), Faculty of Sciences Dhar EL Mehraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco
[2]Laboratory of Engineering Sciences, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
[3]Sumitovant Biopharma, NY, USA

Abstract— **Automatic extraction of the relationship between drugs/chemicals and proteins from biomedical literature is a crucial task for drug discovery and drug-induced adverse reactions. In this paper, we describe our participation in the BioCreative VII Track 1: Text mining drug and chemical-protein interactions (DrugProt). First, we formulate the task as a sequence labelling problem which links the drug/chemical entities to their proteins and interaction types by adopting a unified sequence labelling. Then, we explore a Multi-Task Transfer Learning-based method (MTTL) by training DrugProt task with several clinical and biomedical natural language processing tasks. MTTL adopts a shared representation obtained from the transformer-based models. As final submission, we submitted four runs where the best performing micro average F1 score achieved 0.7133 on the test set. The source code is publically available at https://github.com/drissiya/mttl-drugprot.**

Keywords—Multi-task learning; Transfer learning; Relation extraction; Sequence labelling

I. Introduction

The volume of published biomedical research is expanding at a growing rate. Consequently, it is becoming increasingly challenging for most healthcare researchers to remain up-to-date with the recent knowledge, for instance, the relationship between drugs and chemicals compounds with certain biomedical entities including proteins and genes. Therefore, automatic analysis of biomedical textual data using natural language processing (NLP) techniques is a promising approach. In this context, the critical assessment of information extraction in biology (BioCreative) released several related tasks to promote research efforts for extracting drug-related information from biomedical literature such as protein-protein interaction (PPI) extraction (1), chemical compound and drug name recognition task (CHEMDNER) (2) and text mining chemical-protein interactions (CHEMPROT) (3). Recently, the BioCreative VII organized a shared track called "text mining drug and chemical-protein interactions (DrugProt)". The latter asks participants to develop and assess the current state of the art NLP systems for extracting the relations between chemical /drug compounds and genes/proteins (4).

Traditional approaches for relation extraction task in general and specific domains were primarily focused on machine learning-based methods which exploit various features and feed them into classifiers such as support vector machines (SVM) (5). However, these methods involve heavy handcrafted features which are labor intensive and skill-dependent. Further studies have successfully examined deep neural networks, in particular the combination of recurrent neural networks (RNN) and bidirectional long short-term memory networks (Bi-LSTM) with attention mechanisms (6). However, the performance of these methods heavily depends on the size and quality of labelled training dataset. In light of this, multi-task learning (MTL) (7) is one way to deal with the aforementioned problem and has become one of the more recent core research areas in several NLP tasks. Basic premise behind MTL is learning different tasks together to improve the generalization performance across all tasks. On the other hand, the combination of pre-trained language models (8) based on transformer architecture with MTL have recently started showing promising results on various downstream tasks.

In this paper, we describe our participation in the DrugProt track. We explore the Multi-Task Transfer Learning-based method (MTTL) (9) for extracting the chemical-gene relations. The MTTL adopts a shared representation obtained from the pre-trained language models and ends up with task-specific fine-tuning. We achieved a micro average F1-score of 0.7133 on the test set as our best result.

II. METHODS

A. Pre-processing

We conduct two preprocessing operations. First, we adopt the Punkt sentence detector in NLTK for sentence boundary detection. Then, each sentence is tokenized using WordPiece which splits unseen words into pieces with two hash marks.

B. Formalization of relation extraction

Generally, the relation extraction task is regarded as a classification problem where the instances for each candidate pair are generated from one sentence. For instance, given the example illustrated in Figure 1, there are three instances: the first instance is constructed from the candidate relation between "Progesterone" and "PR", the second instance is generated from the candidate relation between "Progesterone" and "AR", while the third one is constructed from the candidate relation between "Progesterone" and "MR". However, the classification-based methods do not take into account the interaction between relations as they treat them independently.

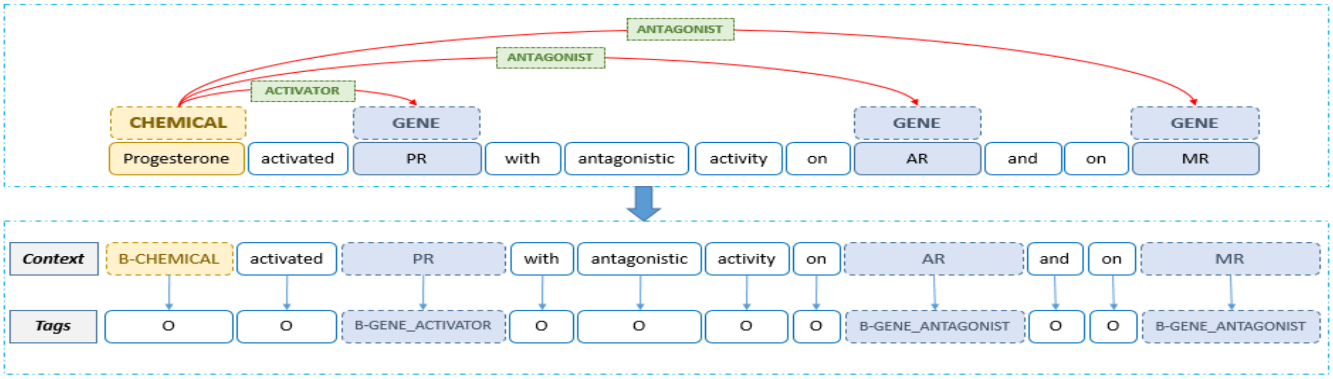

Fig. 1: Example of sequence labelling

To this end, we formulate the relation extraction task as a sequence labelling problem. Specifically, we assume that the chemical entities are given and we need to find their related gene entities and relation types by adopting the BIO segment representation (BIO tags each token as either the beginning (B), inside (I) or outside (O) of the mention). We add the label (X) according to the WordPiece tokenization outcomes. The first piece takes the tag assigned by the original word, while the tag (X) is added for other pieces. To do so, we firstly generate the context related for each chemical entity by replacing its position with a label which is made up of two parts: the boundary tag (B and I) and the CHEMICAL type. This allows the model to benefit from the semantic information of the chemical entity so as to extract its related genes. For instance, in Figure 1, we create a context for "Progesterone" mention by replacing its position with "B- CHEMICAL" in the sentence. Then, we assign a tag to each token of the generated context. In addition to the labels (O) and (X), we define a new label which consists of three parts: the boundary tag of the gene entity, the GENE type and the relation type. Figure 1 illustrates an example of sequence labelling.

C. *Multi-task transfer learning based method*

We explore the MTTL which is based on the MT-DNN architecture (7). It can train several NLP tasks together so as to learn better representations that are helpful for the DrugProt task. Figure 2 shows the overall architecture of MTTL. It contains three main layers: input layer, shared layer and task-specific layer. In the following sections, we will provide the process of each layer.

1) *Input layer*

The first layer takes as input the context $C = \{w_1, \ldots, w_k\}$ of length k. We prune the context to a fixed length n by either trimming longer contexts or padding shorter ones with a special token [PAD]. We set the maximum sequence length to 512 tokens. Let $X = \{x_1, \ldots, x_n\}$ the obtained sequence.

2) *Shared layer*

Each token x_i in X is mapped into contextual embedding vectors. In our experiment, we use the transformer-based models as the shared layer across different tasks. The pre-trained language models include:

- **SciBERT** (10): is pre-trained on full-text papers from Semantic Scholar (18% from the computer science domain and 82% from the biomedical domain) based on BERT architecture. It has two versions of vocabulary: basevocab (the original BERT vocabulary) and scivocab (the vocabulary built using SentencePiece on the scientific corpus). We adopt the scivocab version recommended by the authors.
- **PubmedBERT** (11): is pre-trained on the PubMed articles with a customized vocabulary (built from PubMed articles). There are two models of PubmedBERT: the first model is pre-trained on the abstracts and full text of biomedical literature, while the second one is pre-trained on the abstracts only. We use the first model for evaluation.

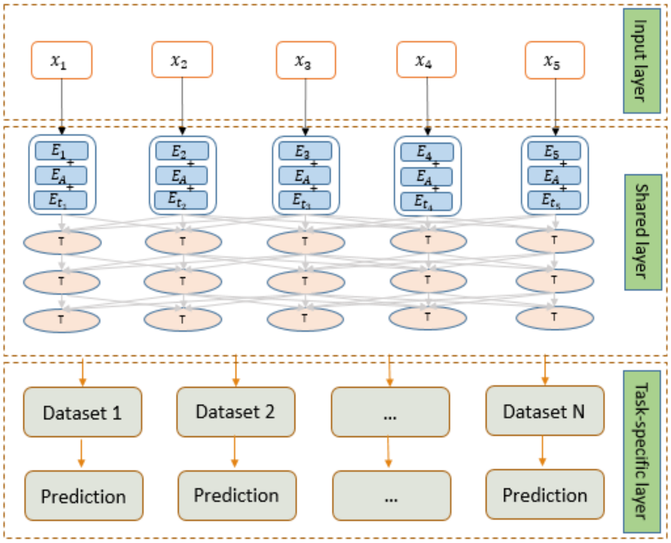

Fig. 2: Overall architecture of MTTL-based method

3) *Task-specific layer*

The last layer uses a fully-connected layer for each task. In our experiments, we applied a logistic regression with softmax function for both sequence labelling and text classification tasks. The integrated tasks are trained using the categorical cross-entropy loss function.

D. *Training*

We apply the mini-batch based stochastic gradient descent to learn the parameters of the model. In each epoch, we randomly select a mini-batch of task t. Then, we update the model by optimizing the task-specific objective for this task.

III. Experimental results

A. Datasets

We used MTTL by training several clinical and biomedical NLP tasks. For relation extraction task as a sequence labelling problem, we applied the following datasets:

TABLE I. STATISTICS OF DRUGPROT DATASET

Relation type	Training	Development
INDIRECT-DOWNREGULATOR	1330	332
INDIRECT-UPREGULATOR	1379	302
DIRECT-REGULATOR	2250	458
ACTIVATOR	1429	246
INHIBITOR	5392	1152
AGONIST	659	131
AGONIST-ACTIVATOR	29	10
AGONIST-INHIBITOR	13	2
ANTAGONIST	972	218
PRODUCT-OF	921	158
SUBSTRATE	2003	495
SUBSTRATE_PRODUCT-OF	25	3
PART-OF	886	258

- **DrugProt** (12): contains 5000 documents from Pubmed titles and abstracts which were split into 3500, 750 and 750 for training, development and test sets. Additional 10000 documents are included as a background set into the test set. The dataset defines 13 relation types. Table 1 shows the statistics for each relation.

- **ChemProt** (3): has 2432 documents from Pubmed titles and abstracts where 1020 of them were used for training set. It contains five different relations.

- **TAC 2017** (13): provides 200 drug labels where 101 of them were used for training set. The dataset defines three relation types: Negated, Hypothetical and Effect.

- **n2c2 2018** (14): contains 505 discharge summaries extracted from MIMIC-III where 303 of them were used for training set. The dataset includes eight relations.

For named entity recognition task, the datasets include:

- **bc2gm** (15): consists of 20000 sentences from PubMed abstracts. It is devoted for extracting gene and alternative gene entities.

- **bc5cdr** (16): contains 1500 Medline abstracts and was used for detecting chemical and disease entities.

- **Ncbi** (17): consists of 793 Medline abstracts. It is annotated with 6892 disease mentions.

For text classification task, we applied the following datasets:

- **i2b2 2010** (18): consists of 426 discharge summaries where 170 documents are used for the training set. It includes eight relations.

- **ADE dataset** (19): contains 2972 Medline case reports with 20967 sentences specifying the presence or absence of adverse drug events (ADE).

B. Evaluation

Once the learned model produces the sequences of the DrugProt task, we extract all the related genes with their relation types for each chemical mention. If the gene mention is correctly identified irrespective of the relation type, then it takes the interactor argument provided in the entities' gold set. We evaluated the performance of the proposed system using the official scripts provided by the task organizers. It adopts the micro-average Precision (P), Recall (R) and F1-score (F1).

C. Implementation details

We adopted the implementation of our MTTLADE system. It is based on PyTorch. We used Adamx optimizer with weight decay, learning rate and batch size being 0.05, 5e-5 and 8, respectively. We used dropout with 0.1 to all the task-specific layer. We set the maximum number of epochs to 8. To avoid the exploding gradient problem, the gradient norm is clipped within 1. We applied the linear learning rate decay schedule with warm-up over 0.1. The source code is publically available at https://github.com/drissiya/mttl-drugprot.

D. Results

We submitted four runs as final submission:

- **Run 1**: MTTL with SciBERT (Datasets: DrugProt, ChemProt, TAC 2017 and n2c2 2018).

- **Run 2**: MTTL with PubMedBERT (Datasets: DrugProt, ChemProt, TAC 2017 and n2c2 2018).

- **Run 3**: MTTL with SciBERT (Datasets: DrugProt, ChemProt, TAC 2017 and n2c2 2018, bc2gm, bc5cdr, ncbi, i2b2 2010 and ADE dataset).

- **Run 4**: MTTL with PubMedBERT (Datasets: DrugProt, ChemProt, TAC 2017 and n2c2 2018, bc2gm, bc5cdr, ncbi, i2b2 2010 and ADE dataset).

Table II shows the system performance of each submitted runs on the test set. It can clearly be seen from the table that the highest performance is achieved by PubMedBERT as the shared layer of MTTL with additional datasets (Run 4). It obtained 0.7569, 0.6744 and 0.7133 for P, R and F1, respectively. The improvement is mainly observed on extracting INDIRECT-UPREGULATOR, ACTIVATOR and INHIBITOR relations with an F1 of 0.7078, 0.7844 and 0.8103, respectively. Compared with other runs, Run 2 performs better on extracting INDIRECT-DOWNREGULATOR, DIRECT-REGULATOR, ANTAGONIST, PRODUCT-OF, SUBSTRATE and PART-OF relations with an F1 of 0.6826, 0.6321, 0.8227, 0.6393, 0.6299 and 0.5894, respectively. Notably, the Run 1 enhances the performance of AGONIST relation against other runs with an F1 of 0.7472. However, all runs fail to correctly extract AGONIST-ACTIVATOR, AGONIST-INHIBITOR and SUBSTRATE_PRODUCT-OF relations. This is due to the fact that the relation distribution is extremely imbalanced where the number of these relations in the training set is very small compared to the other ones as shown in Table I. We carried out an error analysis to gain further insights about our best system (run 4). The main error is caused by missing relations where the GENE mentions are

TABLE II. RESULT OF THE SUBMITTED RUNS ON THE TEST SET

Relation type	Run 1			Run 2			Run 3			Run 4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
INDIRECT-DOWNREGULATOR	0.6454	0.7006	0.6719	0.6656	0.7006	0.6826	0.6845	0.6710	0.6777	0.6845	0.6710	0.6777
INDIRECT-UPREGULATOR	0.6678	0.6606	0.6642	0.7105	0.6823	0.6961	0.6401	0.6678	0.6537	0.7116	0.7039	0.7078
DIRECT-REGULATOR	0.7301	0.4918	0.5877	0.6906	0.5827	0.6321	0.6396	0.4965	0.5590	0.6411	0.5664	0.6014
ACTIVATOR	0.8113	0.6437	0.7178	0.7571	0.7095	0.7326	0.7456	0.6407	0.6892	0.8135	0.7574	0.7844
INHIBITOR	0.8263	0.7649	0.7944	0.8355	0.7830	0.8084	0.7973	0.7488	0.7723	0.8331	0.7887	0.8103
AGONIST	0.8395	0.6732	0.7472	0.7857	0.6534	0.7135	0.7948	0.6138	0.6927	0.8	0.5940	0.6818
AGONIST-ACTIVATOR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AGONIST-INHIBITOR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANTAGONIST	0.8380	0.7777	0.8067	0.8424	0.8039	0.8227	0.8310	0.8039	0.8172	0.8275	0.7843	0.8053
PRODUCT-OF	0.6225	0.5193	0.5662	0.6324	0.6464	0.6393	0.6879	0.5359	0.6024	0.6348	0.6243	0.6295
SUBSTRATE	0.7177	0.4916	0.5835	0.6795	0.5871	0.6299	0.7100	0.4558	0.5552	0.7558	0.5393	0.6295
SUBSTRATE_PRODUCT-OF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PART-OF	0.6886	0.5043	0.5822	0.6923	0.5131	0.5894	0.6428	0.3552	0.4576	0.7115	0.4868	0.5781
Overall	0.7529	0.6383	0.6909	0.7459	0.6822	0.7126	0.7297	0.6180	0.6692	**0.7569**	**0.6744**	**0.7133**

likely linked to the CHEMICAL ones. For instance, given the following sentence: “agomelatine administration protects liver cells from paracetamol-induced hepatotoxicity via antioxidant activity and reduced proinflammatory cytokines, such as TNF-α and IL-6”, the CHEMICAL mention “agomelatine” is related to the GENE mentions “cytokines”, “TNF-α” and “IL-6” with an INDIRECT-DOWNREGULATOR relations. Our system fails to recognize the relation between “agomelatine” and “cytokines”.

IV. CONCLUSION

In this paper, we described our participation in the BioCreative VII DrugProt track. The results demonstrated the effectiveness of the MTTL-based method for extracting the drug-protein relations from biomedical literature. They highlighted that integrating pre-trained language models into multi-task learning is helpful for capturing greater contextualized representation and improving the generalization performance for DrugProt task.

V. REFERENCES

1. M. Krallinger, F. Leitner, C. Rodriguez-Penagos and A. Valencia, "Overview of the protein-protein interaction annotation extraction task of BioCreative II," *Genome Biology,* vol. 9, pp. 1-19, 2008.
2. M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal and A. Valencia, "CHEMDNER: The drugs and chemical names extraction challenge," *Journal of Cheminformatics,* vol. 7, pp. 1-11, 2015.
3. M. Krallinger, O. Rabal and S. A. Akhondi, "Overview of the BioCreative VI chemical-protein interaction Track," in *Proceedings of the sixth BioCreative challenge evaluation workshop2017*, 2017.
4. A. Miranda, F. Mehryary, J. Luoma, S. Pyysalo, A. Valencia and M. Krallinger, "Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations," in *Proceedings of the seventh BioCreative challenge evaluation workshop*, 2021.
5. Y. Peng, A. Rios, R. Kavuluru and Z. Lu, "Extracting chemical-protein relations with ensembles of SVM and deep learning models," *Database,* vol. 2018, 2018.
6. S. Liu, F. Shen, R. K. Elayavilli, Y. Wang, M. Rastegar-Mojarad, V. Chaudhary and H. Liu, "Extracting chemical-protein relations using attention-based neural networks," *Database,* vol. 2018, 2018.
7. X. Liu, P. He, W. Chen and J. Gao, "Multi-Task Deep Neural Networks for Natural Language Understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
8. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *ArXiv,* vol. abs/1810.04805, 2019.
9. E.-d. El-allaly, M. Sarrouti, N. En-Nahnahi and S. O. E. Alaoui, "MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction," *Information Processing & Management,* vol. 58, p. 102473, 2021.
10. I. Beltagy, K. Lo and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.
11. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ArXiv,* vol. abs/2007.15779, 2020.
12. M. Krallinger, O. Rabal, A. Miranda-Escalada and A. Valencia, "DrugProt corpus: Biocreative VII Track 1 - Text mining drug and chemical-protein interactions," 2021.
13. D. Demner-Fushman, S. E. Shooshan, L. Rodriguez, A. R. Aronson, F. Lang, W. Rogers, K. Roberts and J. Tonning, "A dataset of 200 structured product labels annotated for adverse drug reactions," *Scientific Data,* vol. 5, 2018.
14. S. Henry, K. Buchan, M. Filannino, A. Stubbs and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association,* vol. 27, pp. 3-12, 2019.
15. L. Smith, L. K. Tanabe, R. J. n. Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu and Y.-S. Lin, "Overview of BioCreative II gene mention recognition," *Genome biology,* vol. 9, pp. S2-S2, 2008.
16. J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers and Z. Lu, "BioCreative V CDR task corpus: a resource for chemical disease relation extraction," *Database,* vol. 2016, p. baw068, 2016.
17. R. I. Dogan, R. Leaman and Z. Lu, "NCBI disease corpus: a resource for disease name recognition and concept normalization," *Journal of Biomedical Informatics,* vol. 47, pp. 1-10, 2014.
18. O. Uzuner, B. R. South, S. Shen and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association,* vol. 18, pp. 552-556, 2011.
19. H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius and L. Toldo, "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *Journal of Biomedical Informatics,* vol. 45, pp. 1-10, 2012.