Drug-protein relation extraction using ensemble of transformer-based language models

Jenny Copara, Douglas Teodoro
Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
Geneva School of Business Administration, HES-SO University of Applied Sciences and Arts of Western Switzerland, Geneva, Switzerland
Swiss Institute of Bioinformatics, Lausanne, Switzerland

***Abstract*—Drug-protein interactions have become a crucial component to study potential side effects, discover new uses for existing drugs, to name a few applications. We describe our approach based on transformer-based language models to predict relations between chemical and gene entities in DrugProt corpus. Sliding window is used to detect the relation in a passage for the individual models, and then they are combined using majority vote. Our model achieved 60% of F1-score (88% of recall and 45% of precision) in the track 1: text mining drug and chemical-protein interactions at BioCreative VII. Ensemble of transformer-based language models provides a baseline performance for drug-protein interaction extraction.**

Keywords—transformers; relation extraction; ensemble; BERT

I. Introduction

The scientific literature indexed in PubMed during the last four decades has grown exponentially. With more data available, it is increasingly more challenging to take advantage of all the specific knowledge published in these documents. Particularly, drug-related information is of key importance for biology, pharmacological and clinical research. Chemicals and proteins entities offer a wide range of interactions with use cases such as drug discovery, potential side effects, or finding new uses for existing drugs (1).

In the CHEMPROT track at BioCreative VI, chemical-protein interaction has been approached as an ensemble of a support vector machine, a convolutional neural network, and a recurrent neural network achieving 64.10% of F1-score, the best performance in the task (2). Other methods to perform relation extraction were explored by fine-tuning PubMedBERT over wet lab protocols achieving 80.46% (3). Also, in ChEMU-2020, a fine tunning of BioBERT with patent data was performed with 95.36% of F1-score in the event extraction task (4).

The use of contextualized models has been explored, but we can profit from the goodness of different transformer models exploiting the agreement between each model's predictions. Indeed, an agreement approach was explored for named entity recognition for French biomedical entities (8), chemical entities in patent narratives (9), and wet lab protocols entities (10) as ensemble models using a majority voting strategy, achieving robust performance across different corpus and language.

Track 1: Text mining drug and chemical-protein interactions -DrugProt- track (16) at the BioCreative VII workshop proposes a challenge to automatically detect relations between chemical compounds/drugs and genes/proteins. In this work, we explored ensembles of transformer-based language model to predict relation types in DrugProt using a majority vote strategy. We describe our methodology and show how individual fine-tuned models compare to different ensemble setups for the chemical-protein relations assessed in the shared task.

II. Methods

We explore the effectiveness of transformer-based language models to extract relations between chemical and gene/protein entities. Predictions of the individual models are used to create ensemble models with a majority voting criterion. The individual models assessed in our experiments are: BERT-base, BERT-large, SciBERT, PubMedBERT, and Biomed-RoBERTa. Evaluation is done using micro precision, micro recall, and micro F1-measure through the DrugProt evaluation library.

A. Transformers for relation extraction

We select five transformer models based on content coverage and tokenizer family. Bidirectional Encoder Representations from Transformers (BERT)-base-cased and BERT-large-cased are trained on large English corpora extracted from BookCorpus and Wikipedia (5). Base and large models differ in the number of attention heads, i.e., 12 and 16 respectively. SciBERT is a BERT model trained on scientific texts taken from Semantic Scholar (6). Full text of manuscripts was considered for training. The subdomains covered in the SciBERT corpus are 18% of computer science papers and 82% of the biomedical domain. PubMedBERT is a BERT version trained from scratch on a collection of 14 million PubMed abstracts (7). The specialized vocabulary offers a more robust language model in biomedical natural language processing tasks. Finally, Biomed-RoBERTa is a language model based on RoBERTa (13) and trained on the Semantic Scholar Open Research Corpus (15). This model covers mostly papers in the field of medicine, biology, and physics (14).

The transformer-based language models are fine-tuned in the DrugProt data provided during the challenge. The fine-tunning is done with a maximum sequence length of 512 tokens, batch size of 32, and AdamW as optimizer. We follow

the suggestion of BERT authors (7) about the number of epochs (4) and learning rate (5e-5). Tokenization is driven by the original model's tokenizer, i.e., BERT based models use WordPiece (11) while RoBERTa based model uses Byte-Pair-Encoding (12). Implementations are based on the SimpleTransformer library, and models loaded from Huggingface.

B. *Ensemble models*

We submit three runs containing ensemble models. The strategy to get our ensembles is based on a majority vote, where each model produce their predictions independently (8-10). This process is illustrated in Fig. 1. For a given pair of entities, each model predicts a relation type. Each prediction acts as a voting individual. To assign a prediction to a pair of entities, the candidate relation type should achieve a majority vote. In run-2, named as 'ensemble 1', we combined BERT-base, SciBERT, and PubMedBERT. In run-3, named as 'ensemble 2', we used SciBERT, PubMedBERT, and Biomed-RoBERTa. In run-4, named as 'ensemble 3', we combined the five models.

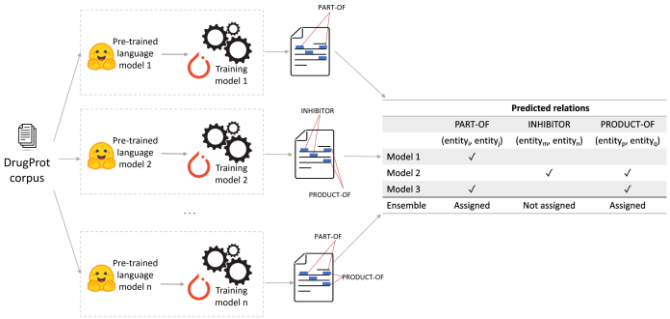

Fig. 1. Majority vote strategy applied in ensemble models

C. *Dataset and pre-processing*

In the BioCreative VII-Track 1: DrugProt, an annotated dataset of PubMed abstracts with chemical-protein relations using the BRAT standoff format was provided (16). Abstracts were annotated with entities that might have a relation. In this dataset, a relation comprises always two entities and the following relation types were provided: indirect-downregulator, indirect-upregulator, direct-regulator, activator, inhibitor, agonist, antagonist, agonist-activator, agonist-inhibitor, product-of, substrate, substrate_product-of, and part-of. The dataset was organized in train and development sets, with 3,500 and 750 abstracts respectively.

Relation extraction is considered a text classification task. We generated passages of text associated with a relation type. First, we processed the abstracts to generate the input for our models. Fig. 2 shows this process. For each combinatorial pair of entities in the abstract, we replaced their passages with the masks 'ENTITY_1' and 'ENTITY_2', according to the order in the annotated relations (ENTITY_1 is a ... and ENTITY_2 is a ...). Then, we associated the masked pairs to their respective relation (including "no relation"). Next, the abstracts were divided into sentences using the spaCy library. Finally, we applied a sliding window of size k (k is the number of sentences) over the sentences to generate positive and negative samples of relation types. This process ended up with a set of passages of size k-sentences associated with the corresponding relation type for the masked entities, see the right column in Fig. 2. We omit the passages that do not contain both entity masks as a relationship needs two entities. During training, we found that a window of 3 sentences gets the best performance.

Our models needed to recognize also when a pair of entities are not related. During training, we found that the number of negative samples was too high as most of entity pairs created combinatorially do not have a relation, generating unbalanced data. Thus, we balanced the generation of negative samples by the number of positive samples in the abstract.

III. Results and Discussion

A. *Official results in main track DrugProt*

Our team - DigiLab-UG - submitted four runs for the main track in DrugProt. The test set of the main track contains 10,750 abstracts annotated with entities. Our run-1 is based on the BERT-base model and is taken as a baseline. In run-2 (BERT only ensemble), we used an ensemble of BERT-base, SciBERT, and PubMedBERT. In run-3 (specialised ensemble), we used an ensemble of SciBERT, PubMedBERT, and Biomed-RoBERTa. In run-4 (all ensemble), we used an ensemble of the five transformer models. After the prediction of each of our models, we keep only relations where the first entity is of chemical type and the second is gene type.

The detailed performance of each run can be seen in Table I. Recall for all our runs is significantly higher than the mean recall in DrugProt, which is 62.91%. Concerning the mean of precision and F1-score, we are under the mean. Run-4 overall provides our best performance with 59.59% of F1-score, thanks particularly to its precision. However, run-3 achieves the best recall with 88.05%. All ensembles outperform the baseline model.

TABLE I. Official results of our submission

ID	Name	Precision	Recall	F1-score
run-1	BERT-base-cased	0.3922	0.8564	0.5380
run-2	Ensemble1	0.4344	0.8693	0.5793
run-3	Ensemble2	0.4391	**0.8805**	0.5860
run-4	Ensemble3	**0.4507**	0.8794	**0.5959**

Table II shows the performance of our runs by relation type. Run-1, i.e., BERT-base, has a higher performance for the agonist, antagonist and inhibitor relations (all above 60% F1-score). Runs 2 to 4, i.e., the ensemble models, identify better antagonist and agonist-inhibitor relations (above 70% and 80% F1-score, respectively). The ensemble also outperforms the baseline for the relation types.

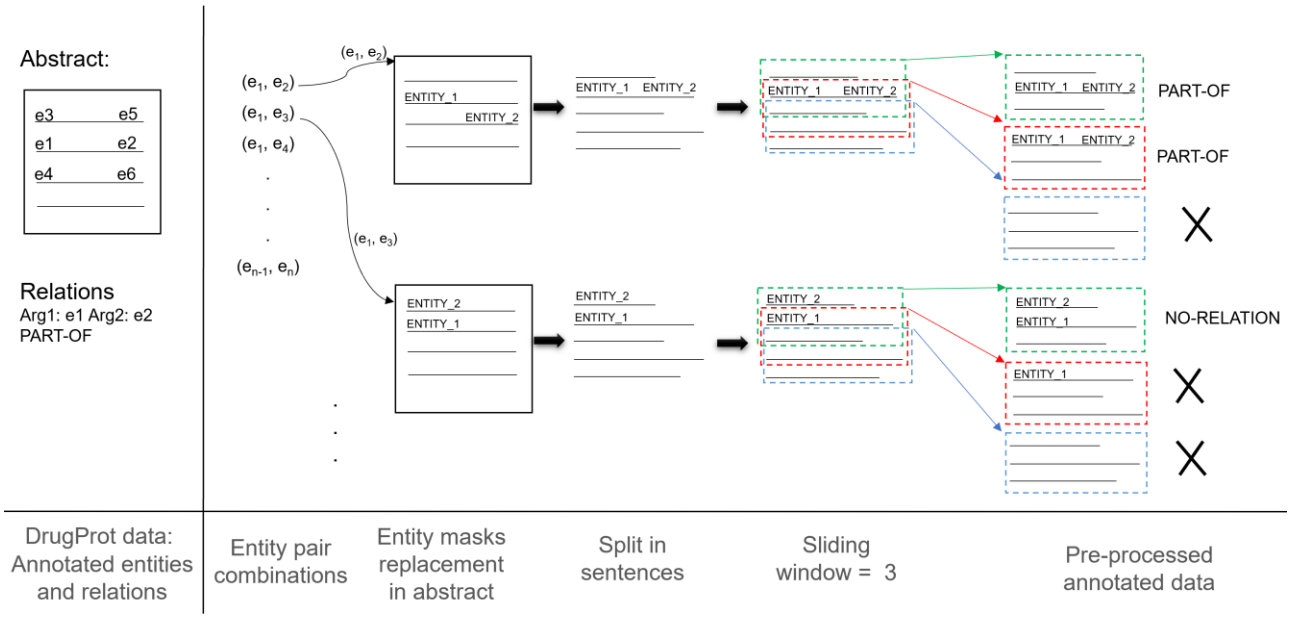

Fig. 2. Pre-processing process to get data for our models

TABLE II. OFFICIAL F1-SCORE BY RELATION TYPE OF OUR RUNS

Relation Type	Run-1	Run-2	Run-3	Run-4
Activator	0.5854	0.5998	0.6083	**0.6318**
Agonist	0.6090	0.6382	0.6323	**0.6712**
Agonist-activator	0.0000	0.0000	0.0000	0.0000
Agonist-inhibitor	0.4615	**0.8571**	**0.8571**	**0.8571**
Antagonist	**0.6475**	0.7084	**0.7327**	0.7202
Direct-regulator	0.4885	0.5003	0.4969	**0.5204**
Indirect-downregulator	0.5541	0.6296	**0.6366**	0.6286
Indirect-upregulator	0.5187	0.5819	0.5964	**0.5998**
Inhibitor	0.6437	0.6793	0.6828	**0.6922**
Part-of	0.4447	0.5424	0.5674	**0.5764**
Product-of	0.3406	0.4028	**0.4183**	0.4164
Substrate	0.4603	0.4802	0.4846	**0.4913**
Substrate_product-of	0.0000	0.0000	0.0000	0.0000

None of our models were able to identify agonist-activator and substrate_product-of relations in the test set. We analyze this result in the development set and found that agonist-activator is confused with activator relation. We also observe that activator relation has a much higher proportion of annotations than agonist-activator, i.e., 8.27% and 0.17% respectively, in the train set. We believe this could be the reason for this confusion. We also noticed confusion between substrate_product-of and substrate relations. The proportion of relations in the train set are 11.59% for substrate and 0.14% for substrate_product-of. However, our models can predict these two relations using the training data. In other words, our models have learned those relations but they are not able yet to predict accurately them.

B. *Experiments in development set*

As the official test set is not publicly available, to further analyze our results we evaluated the performance in the development set (also as out-of-sample). Table III shows the micro precision, recall, and F1-score for all the explored transformer-based language models and ensembles.

TABLE III. PERFORMANCE OF EXPLORED MODELS EVALUATED IN DEVELOPMENT SET

Model	Precision	Recall	F1-score
BERT-base	0.8750	0.8470	0.8610
BERT-large	0.8840	0.8550	0.8690
SciBERT	0.8790	0.8600	0.8700
PubMedBERT	0.8870	0.8730	0.8800
Biomed_RoBERTa	0.8860	0.8730	0.8790
Ensemble1	0.8990	0.8690	0.8840
Ensemble2	0.8990	0.8730	0.8860
Ensemble3	**0.9050**	**0.8750**	**0.8900**

In Table III, the ensemble 3 (equivalent to run-4) model achieves the highest performance, which is similar to the official results. However, we notice a significant drop in performance in the official test set (from 89% to 60% F1-score). The best ensemble model outperforms the baseline in terms of F1-score by almost 3 percentage point. We can also see in Table III how the specialized transformer-based language models have better performance compared to the models trained in a general corpus (BERT-base and -large).

Similar to Table II, Table IV shows the F1 measure achieved in the development set for the different relation types.

Differently from the official results, the agonist-activator is now detected. The models are still unable to detect the substrate_product-of class. Surprisingly, agonist-inhibitor relation is always classified correctly in the development set for the models used in the official submission. However, this is not the case for SciBERT and BERT-large (93.7% and 93.6%, respectively).

As described in Section II, subsection C, during the pre-processing stage we ended up with several passages associated with the same relation type, as a result of the sliding window process. In official submission, we assigned only the last predicted class for a given entity-pair when several passages belong to the same pair of entities. We modified this algorithm to choose the relation type with the highest prediction score for the same entity-pair among all the possibilities. Tables III and IV contain this modification.

Among the five transformers, PubMedBERT is the model with the best performance in the development set, achieving 88.7% of precision, 87.3% of recall, and 88.0% of F1-score (results not shown). This could be due to the fact that PubMedBERT was trained with PubMed abstracts, which is the same source of the content of the annotated abstracts provided in this track.

TABLE IV. F1-SCORE BY RELATION TYPE OF EXPLORED MODELS EVALUATED IN DEVELOPMENT SET

Relation Type	BERT-base	Ensemble 1	Ensemble 2	Ensemble 3
Activator	0.8140	0.8540	**0.8610**	0.8570
Agonist	0.8480	0.8660	0.8640	**0.8710**
Agonist-activator	0.1540	**0.5710**	**0.5710**	**0.5710**
Agonist-inhibitor	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Antagonist	0.9310	**0.9500**	0.9450	0.9450
Direct-regulator	0.8000	**0.8380**	0.8340	**0.8380**
Indirect-downregulator	0.8100	0.8290	0.8380	**0.8490**
Indirect-upregulator	0.8470	0.8710	0.8840	**0.8860**
Inhibitor	0.9250	**0.9310**	0.9260	**0.9310**
Part-of	0.8460	0.8940	**0.8970**	0.8830
Product-of	0.7160	0.7600	0.7890	**0.7940**
Substrate	0.8650	0.8850	0.8910	**0.9020**
Substrate_product-of	0.0000	0.0000	0.0000	0.0000

In Table V, we provide an example of a sliding window for the relation agonist in abstract 16324695 taken from the development set. The window size is 3 sentences. The passages are created when argument 1 of the relation is the entity T7, i.e., isoprenaline, and argument 2 is entity T20, i.e., beta(2)-adrenoceptor. The sliding window passage contains the entity masks explained in the pre-processing subsection of Section II, C.

TABLE V. SLIDING WINDOW EXAMPLE FOR THE AGONIST RELATION IN ABSTRACT WITH PMID 16324695 WHEN ARG1 IS ENTITY T7 AND ARG2 IS ENTITY T20

Window ID	Sliding window passage
1	Protein kinase C potentiates homologous desensitization of the beta2-adrenoceptor in bovine tracheal smooth muscle. Preincubation (30 min) of bovine tracheal smooth muscle with various concentrations (0.1, 1 and 10 microM) of fenoterol decreased ENTITY_1-induced maximal relaxation (E(max)) of methacholine-contracted preparations in a concentration dependent fashion, indicating desensitization of the ENTITY_2. Preincubation with 1 microM of the protein kinase C (PKC) activator phorbol 12-myristate 13-acetate (PMA) caused a small but significant decrease in isoprenaline-induced E(max), indicating activated PKC-mediated heterologous beta(2)-adrenoceptor desensitization.
2	Preincubation (30 min) of bovine tracheal smooth muscle with various concentrations (0.1, 1 and 10 microM) of fenoterol decreased ENTITY_1-induced maximal relaxation (E(max)) of methacholine-contracted preparations in a concentration dependent fashion, indicating desensitization of the ENTITY_2. Preincubation with 1 microM of the protein kinase C (PKC) activator phorbol 12-myristate 13-acetate (PMA) caused a small but significant decrease in isoprenaline-induced E(max), indicating activated PKC-mediated heterologous beta(2)-adrenoceptor desensitization. To investigate the capacity of activated PKC to regulate homologous desensitization, we incubated the smooth muscle strips with the combination of both 1 microM PMA and 1 microM fenoterol.

In Table VI, we also show the predictions of the explored models over the sliding windows shown in Table V. Relation types in bold are chosen to be assigned as relation using the criterion of the highest prediction score. After we get predictions of all our models we generate the ensemble models.

TABLE VI. PREDICTION OF SLIDING WINDOWS FOR A POSITIVE SAMPLE

Window ID	BERT-base	SciBERT	PubMed BERT	Biomed-RoBERTa	BERT-large
1	No-Relation	**Activator**	**Agonist**	Agonist	Agonist
2	**No-Relation**	Agonist	Agonist	**Agonist**	**Agonist**

Table V shows a positive example of relation extraction in the DrugProt track, i.e., agonist relation. In Table VII can be seen an example when the class is no-relation. Here the first argument of the relation is entity T2, i.e., fenoterol, while the second argument is entity T7, i.e., isoprenaline. Entity masks are replaced in the passage, similar to Table V.

TABLE VII. SLIDING WINDOW EXAMPLE FOR A NO-RELATION TYPE IN ABSTRACT WITH PMID 16324695 AND ENTITIES T2 AND T7

Window ID	Sliding window passage
1	Protein kinase C potentiates homologous desensitization of the beta2-adrenoceptor in bovine tracheal smooth muscle. Preincubation (30 min) of bovine tracheal smooth muscle with various concentrations (0.1, 1 and 10 microM) of ENTITY_1 decreased ENTITY_2-induced maximal relaxation (E(max)) of methacholine-contracted preparations in a concentration dependent fashion, indicating desensitization of the beta(2)-adrenoceptor. Preincubation with 1 microM of the protein kinase C (PKC) activator phorbol 12-myristate 13-acetate (PMA) caused a small but significant decrease in isoprenaline-induced E(max), indicating activated PKC-mediated heterologous beta(2)-

Window ID	Sliding window passage
	adrenoceptor desensitization.
2	Preincubation (30 min) of bovine tracheal smooth muscle with various concentrations (0.1, 1 and 10 microM) of ENTITY_1 decreased ENTITY_2-induced maximal relaxation (E(max)) of methacholine-contracted preparations in a concentration dependent fashion, indicating desensitization of the beta(2)-adrenoceptor. Preincubation with 1 microM of the protein kinase C (PKC) activator phorbol 12-myristate 13-acetate (PMA) caused a small but significant decrease in isoprenaline-induced E(max), indicating activated PKC-mediated heterologous beta(2)-adrenoceptor desensitization. To investigate the capacity of activated PKC to regulate homologous desensitization, we incubated the smooth muscle strips with the combination of both 1 microM PMA and 1 microM fenoterol.

In Table VIII is shown the prediction of our models for these passages. In this example, most of the models classify the passages as an inhibitor. Only SciBERT and Biomed-RoBERTa predict the correct class.

TABLE VIII. PREDICTION OF SLIDING WINDOWS FOR A NEGATIVE SAMPLE

Window ID	BERT -base	SciBERT	PubMed BERT	Biomed-RoBERTa	BERT -large
1	**Inhibitor**	**No-Relation**	Inhibitor	**No-Relation**	**Inhibitor**
2	Inhibitor	Inhibitor	**Inhibitor**	No-Relation	Inhibitor

IV. CONCLUSION

We presented our approach for drug-protein relation extraction using ensembles of transformer-based language models based on a majority voting strategy. We compared the results of the individual models with the ensembles and assessed the performance for the different relation types. The combination of individual models adds an important contribution to the performance. However, for sub-represented classes in the training set, results are still poor. The methodology provides a baseline approach for extracting drug-protein relations, being close to the average models in the BioCreative VII DrugProt challenge. In future work, we will explore strategies to overcome the 0-performance in substrate_product-of and agonist-activator.

ACKNOWLEDGMENT

Funding for this work is provided by the CINECA project (H2020 No 825775) and Innosuisse project funding number 46966.1 IP-ICT.

REFERENCES

1. Tsubaki, M., Tomii, K., & Sese, J. (2019). Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics (Oxford, England)*, *35*(2), 309–318.

2. Peng, Y., Rios, A., Kavuluru, R., & Lu, Z. (2018). Extracting chemical-protein relations with ensembles of SVM and deep learning models. *Database: The Journal of Biological Databases and Curation*, *2018*. https://doi.org/10.1093/database/bay073.

3. Tabassum, J., Xu, W., & Ritter, A. (2020). WNUT-2020 task 1 overview: Extracting entities and relations from wet lab protocols. *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*.

4. He, J., Nguyen, D.Q., Akhondi, S.A., Druckenbrodt, C., Thorne, C., Hoessel, R., Afzal, Z., Zhai, Z., Fang, B., Yoshikawa, H., Albahem, A., Wang, J., Ren, Y., Zhang, Z., Zhang, Y., Hoang Dao, M., Ruas, P., Lamurias, A., M. Couto, F., Copara, J., Naderi, N., Knafou, J., Ruch, P., Teodoro, D., Lowe, D., Mayfield, J., Koksal, A., Donmez, H., Ozkırımlı, E., Ozgur, A., Mahendran, D., Gurdin, G., Lewinski, N., Tang, C., T.McInnes, Bridget C.S., M., RK Rao., P., Lalitha Devi, S., Cavedon, L., Cohn, T., Baldwin, T., Verspoor, K. (2020) . An extended overview of the CLEF 2020 ChEMU Lab : information extraction of chemical reactions from patents. In Proceedings of the CLEF 2020 conference.

5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1423.

6. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1371.

7. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. In *arXiv [cs.CL]*. https://doi.org/10.1145/3458754.

8. Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P., & Teodoro, D. (n.d.). *Contextualized french language models for biomedical named entity recognition*. Loria.Fr. Retrieved October 7, 2021, from https://jep-taln2020.loria.fr/wp-content/uploads/JEP-TALN-RECITAL-2020_paper_215.pdf.

9. Copara, J., Naderi, N., Knafou, J., Ruch, P., & Teodoro, D. (2020). Named entity recognition in chemical patents using ensemble of contextual language models. http://ceur-ws.org/Vol-2696/paper_219.pdf.

10. Knafou, J., Naderi, N., Copara, J., Teodoro, D., & Ruch, P. (2020). BiTeM at WNUT 2020 shared task-1: Named entity recognition over wet lab protocols using an ensemble of contextual language models. *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*, 305–313.

11. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation system: Bridging the gap between human and Machine Translation. In *arXiv [cs.CL]*. http://arxiv.org/abs/1609.08144.

12. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv [cs.CL]*. http://arxiv.org/abs/1907.11692.

14. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *arXiv [cs.CL]*. http://arxiv.org/abs/2004.10964.

15. Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2019). S2ORC: The semantic scholar open research corpus. In *arXiv [cs.CL]*. http://arxiv.org/abs/1911.02782.

16. Miranda, A., Mehryary, F., Luoma, J., Pyysalo, S., Valencia, A. and Krallinger, M. (2021). Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. Proceedings of the seventh BioCreative challenge evaluation works.