Recognizing Chemical Entity in Biomedical Literature using a BERT-based Ensemble Learning Methods for the BioCreative 2021 NLM-Chem Track

Yu-Wen Chiu[1], Wen-Chao Yeh[2], Sheng-Jie Lin[1], Yung-Chun Chang[1,*]
[1, *]Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan
[2]Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan

***Abstract*—It is challenging and time-consuming to extract crucial information from biomedical literature, which is often unstructured and noisy. To reduce human effort and increase efficiency, many organizations have invested resources in the research of Natural Language Processing (NLP) technology. In light of this, there are two tasks in the BioCreative VII Track 2 for creating effective NLP downstream applications, namely, Chemical Mention Recognition and Chemical Normalization. In this paper, we propose different ensemble learning methods to integrate multiple BERT pre-trained models for recognizing chemical entities. A dynamic programming-based method is adopted to perform entity linking for normalization. For the Chemical Mention Recognition, we achieved F_1-scores of 85% (strict) and 91.8% (approximate). Moreover, our model performs very well on the task of Chemical Normalization, with strict F_1-score of 80.7% and approximate F1-score of 80.2%. Our systems outperform the benchmark and median significantly, and achieve remarkable performances.**

Keywords—chemical named entity recognition and normalization; BERT; ensemble learning

I. INTRODUCTION

Text data is always noisy, and it is time-consuming to collect useful information from a wide variety of sources. To reduce human effort and efficiently obtain information, many organizations have applied Natural Language Processing (NLP) technology for process automation. Information retrieval (IR) and information extraction (IE) are becoming increasingly popular in the biomedical area. Text mining is another method which has been applied to biomedical research in order to boost the number of scientific publications collected. Computational approaches such as NLP, for digesting biomedical literature could help biologists, bioinformaticians, and database curators get faster access to important textual material. In chemical domains, such as pharmaceuticals and academic publications, chemical NLP and text mining technologies (ChemNLP or chemical text mining) enable access to and integration of information from unstructured data. This also encouraged the cooperation of the biomedical community and text mining community. For example, finding literature in PubMed relevant to a user's information requirement is not always straightforward (1). Hence, IE techniques will play a key part in making the task more feasible. To perform IE and IR in the biomedical field, it is first required to conduct named entity recognition. Different forms of chemical terms and errors in recognizing relevant biological entities are a significant barrier of context comprehension, overall retrieval, and classification. Therefore, automated identification methods are presented in several journals to improve downstream biomedical NLP tasks. As an example, NLP has been used for gene-protein interaction and protein-protein interaction annotation of full-text articles (2). Several challenges have been addressed in earlier works on biological named entity recognition (NER) and chemical normalization, for instance, CHEMDNER and CDR tasks at previous BioCreative workshops (3-4). In recent years, more advanced IR and IE are required for the purpose of extracting fine-grained information, for example, chemical compound structures, different biological reactions between disease, gene and other chemicals, etc.

In light of this, two tasks are included in the BioCreative VII Track 2 to encourage a more effective NLP application (5-6). First, teams are required to predict all mentioned chemicals in the full-text article and normalize them into a canonical form. Next, teams are to predict which compound should be indexed from newly published full-text articles. To recognize chemical mentions, we try to integrate different BERT models through ensemble learning. A dynamic programming-based method is developed to further conduct entity linking for chemical normalization. The results demonstrate that the proposed system can outperform the benchmark and median of comparisons, as well as achieve remarkable performances.

II. DATASET AND METHODS

A. Dataset

This research adopts the NLM-Chem corpus as the training set, which is the largest corpus of full-text articles annotated with chemical entities at a fine granularity (7). The NLM Chemical corpus contains a total of 38,342 manual chemical mention annotations corresponding to 4,867 unique chemical name annotations, and normalized to 2,064 Medical Subject Headings (MeSH) identifiers. They are extracted from 150 exhaustively examined full-text articles. The training dataset also incorporated CHEMDNER and BCCDR, which include about 11,500 PubMed abstracts as additional data. On the other hand, the full text of 1000+ PubMed articles scheduled for human indexing in 2021 is distributed as the test set. NLM expert who annotated the training set will fully annotate a subset of these articles for all occurrences of chemical mentions as the gold standard for the Chemical Identification task. The

TABLE I. BIOCREATIVE VII NLM-CHEM TRACK PERFORMANCE

Submission	**Chemical Mention Recognition**		**Chemical Normalization to MeSH IDs**	
	Strict	Approximate	Strict	Approximate
	Precision / Recall / F_1-score			
Subtask1-Run-1	0.8616 / 0.8415 / 0.8515	0.9293 / 0.9028 / 0.9158	0.7874 / 0.8281 / **0.8072**	**0.7530** / 0.8643 / **0.8015**
Subtask1-Run-2	0.8461 **/ 0.8583 / 0.8521**	0.9152 **/ 0.9215 / 0.9183**	0.7748 / **0.8315** / 0.8021	0.7341 / **0.8669** / 0.7914
Subtask1-Run-3	0.8580 / 0.8409 / 0.8494	0.9257 / 0.9045 / 0.9149	**0.7876** / 0.8272 / 0.8069	0.7462 / 0.8606 / 0.7959
Subtask1-Run-4	0.8345 / 0.7374 / 0.7830	0.9123 / 0.7993 / 0.8521	0.7571 / 0.7886 / 0.7725	0.7311 / 0.8441 / 0.7774
Subtask1-Run-5	**0.8618** / 0.8209 / 0.8409	**0.9303** / 0.8822 / 0.9056	0.7821 / 0.8226 / 0.8019	0.7468 / 0.8569 / 0.7936

human expert indexing of all articles will be the gold-standard for the Chemical Indexing task. In order to scale up the dataset, we integrate CHEMDNER corpus and the CDR corpus from previous BioCreative tasks.

B. Methods

In the Chemical Mention Recognition and Chemical Normalization task, the first sub-track of BioCreative VII Track 2, the official script provides us with the baseline calculation code from the beginning. We first applied greedy search to test hundreds of BERT models trained by biomedical text. Moreover, we also carry out distinctly different model architecture to find the better one for more flexible adjustment of basic environment settings or hyperparameters. The different systems we developed for this track are described below (Here, we use submission id as the name of a method):

- *Subtask1-Run-1:* in order to find out which BERT pre-trained model is suitable for this task, we applied hundreds of biomedical-related BERT pre-trained models to find out how well biological language models perform. We conduct experiments to evaluate the performance using the NLM-Chem corpus under 10-fold cross-validation. Multiple BERT models proposed within the last three years with different settings are utilized in these experiments. The results indicate that the Sultan model (8) significantly improves the performance of recognizing chemical entities. This pretrained BERT model is built on top of the ELECTRA architecture (9) and trained with PubMed abstracts and a biomedical domain vocabulary for 434K steps with a batch size of 4096. In this submission, we adopt NLM-Chem corpus for Sultan model training.

- *Subtask1-Run-2:* according to the results in the experiments of the first submission, the performance of PubMedBERT (10) is comparable with Sultan. In order to gain the advantages of different BERT models, an ensemble model was adopted in this submission. We integrate the prediction results from Sultan, PubMedBERT, and PubMedBERT fine-tuned versions through a majority voting mechanism.

- *Subtask1-Run-3:* we integrate different batch size settings (i.e., 8 and 16) of Sultan models using ensemble learning with majority voting mechanism. In addition, we add positional information of text into models using section type of the full text provided by the NLM-Chem dataset, including abstract, introduction, Table, etc.

- *Subtask1-Run-4:* in this submission, we investigate the improvement from a larger training dataset. Hence, we further merge NLM-Chem, CHEMDNER, and CDR corpora to scale up the training instances, and retrain the Sultan model used in the first submission.

- *Subtask1-Run-5:* we explore the possibility of improving the model in submission 2 through a larger dataset. Therefore, we exploit the merged dataset in *Subtask1-Run-4* to retrain the model used in *Subtask1-Run-2* for this submission.

For the Chemical Normalization task, we adopt a dynamic programming-based method to filter predicted identification and their Mesh ID. First, we extract all existing dataset's identification and MeSH ID as a knowledge base. Next, all predicted terms are looked up in this knowledge base in order to obtain the pair mapping. If none is found, the Edit Distance method with 90% similar parameters will be applied to search the MeSH code book and retrieve the most similar identification term and their mapping ID. An empty value will be assigned if the predicted identification cannot find any mapping MeSH ID from the above steps. Parallel programming is adopted to save search time for a high volume of identification terms.

III. RESULT AND DISCUSSION

A. Results

The models were implemented with PyTorch, a Python deep learning library. BERT pretrained models adopted from HuggingFace package include BioM-ELECTRA-Large-Discriminator for Sultan model, BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext-finetuned-BC2GM and BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext for PubMedBERT and its fine-tuned version. In this competition, precision, recall, and F_1-score are adopted for evaluating performance for chemical entity recognition. Two kinds of grading rules are utilized for both tasks in competition. For Chemical Mention Recognition, *Strict* requires that the start and end offsets match exactly, while the *Approximate* only requires

that they overlap. For Chemical Normalization to MeSH IDs evaluation, *Strict* compares the set of identifiers directly, while the *Approximate* first augments the sets of identifiers with (a subset of) parent identifiers as described for LCAF evaluation (11-12).

Table I shows the performance of our submissions in BioCreatve VII NLM-Chem Track. The performances of our five submissions to the Chemical Mention Recognition task range from 78~86% and 85~92% F_1-score on aspects of Strict and Approximate modes, respectively. It is interesting to note that the performance of submission 4 is inferior compared to others. The results suggest that those models trained with the large dataset performed slightly worse in the Chemical Mention Recognition part. We can observe this phenomenon from Subtask1-Run-5. Because the content of the specific dataset for this competition is the full text of research articles, and the corpora of the other two previous datasets are mainly abstracts. The abstract consists of relatively concise and clear text, which may encourage the model to focus on the compendious text too much. Hence, the prediction performance of the full text is reduced. Notably, Subtask1-Run-2 method can achieve the best performance with 85.21% and 91.83% F_1-score on aspects of Strict and Approximate modes, respectively. This illustrates the key benefit of using ensembles is to improve the average prediction performance over any contributing member in the ensemble. The mechanism for improved performance with ensembles is often the reduction in the variance component of prediction errors made by the contributing models.

For the Chemical Normalization, our prediction performance of five submissions can achieve around 77~80% and 78~80% F_1-score on aspects of Strict and Approximate, respectively. The results display that the strict and approximate scores not being much different. This may because the length of the token is short on average, which causes using the edit distance to partially match token sequences to be efficient in searching out correct answers in the MeSH hierarchy. Note that in *Subtask1-Run-2*, the BERT-based ensemble learning method can achieve the best performance on recognizing chemical entities. However, based on the result of this method will slightly decrease the performance on the Chemical Normalization task. This is because the precision of Chemical Mention Recognition in *Subtask1-Run-1* is higher than *Subtask1-Run-2*. Based on the chemical entity recognition results of *Subtask1-Run-1*, the edit distance method can normalize chemical entities more accurately, and thus achieves the best performance with 80.72 and 80.15 F_1-score on aspects of Strict and Approximate, respectively.

IV. Conclussion

When chemicals are mentioned in PubMed articles, the description of a chemical substance is often related to the position that they appear. Even if they are the same compound, there will be different expressions due to the context in which they appear. If all relevant compounds in each document are found and organized into structured data, future scholars or students can use this method when studying the effects of chemicals on different diseases and different biomedical topics. Structured relational databases can also be used to find the chemical substance of interest. In this competition, we proposed five different models for recognizing chemical entities in the full-text research articles, and further used the dynamic programming method to normalize them into a canonical form. The results provided by the organizers show that our proposed methods can achieve a remarkable performance with about 80% F_1-score on the complete chemical identification process.

Acknowledgment

This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 109-2410-H-038 -012 -MY2.

References

1. Islamaj Dogan, R., Murray, G. C., Névéol, A., & Lu, Z. (2009). Understanding PubMed® user search behavior through log analysis. *Database*, 2009.

2. Islamaj Doğan, R., Kim, S., Chatr-Aryamontri, A., Chang, C. S., Oughtred, R., Rust, J., ... & Tyers, M. (2017). The BioC-BioGRID corpus: full text articles annotated for curation of protein–protein and genetic interactions. *Database*, 2017.

3. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *J. Cheminf.*, 7(1), 1-11.

4. Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., ... & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

5. Leaman, R., Islamaj, R., Lu, Z. (2021) Overview of the NLM-Chem BioCreative VII track: Full-text Chemical Identification and Indexing in PubMed articles. *Proceedings of the seventh BioCreative challenge evaluation workshop*.

6. Islamaj, R., Leaman, R., Cissel, D., …& Lu, Z. (2021) The chemical corpus of the NLM-Chem BioCreative VII track: Full-text Chemical Identification and Indexing in PubMed articles. *Proceedings of the seventh BioCreative challenge evaluation workshop*

7. Islamaj, R., Leaman, R., Kim, S., Kwon, D., Wei, C. H., Comeau, D. C., ... & Lu, Z. (2021). NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data*, 8(1), 1-12.

8. Alrowili, S., & Vijay-Shanker, K. (2021, June). BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 221-227).

9. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

10. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare.*

11. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., ... & Paliouras, G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinf.*, 16(1), 1-28.

12. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., & Androutsopoulos, I. (2015). Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Min. Knowl. Discov.*, 29(3), 820-865.