Fine-tuning transformers for automatic chemical entity identification in PubMed articles

Robert Bevan and Matthew Hodgskiss
Medicines Discovery Catapult, Macclesfield, United Kingdom

***Abstract*— This systems description paper details our entry to the BioCreative 7 NLM-Chem track challenge. We compared two different approaches to chemical entity recognition. First, we fine-tuned the PubMedBERT transformer model using the NLM-Chem dataset only. We then tried building a stacking model using the outputs from the NLM-Chem PubMedBERT transformer and two additional PubMedBERT transformers: one fine-tuned with BioCreative IV's CHEMDNER dataset, and the other fine-tuned with BioCreative V's CDR dataset. We observed no difference in performance between the single PubMedBERT transformer, fine-tuned using the NLM-Chem dataset only, and the stacking model. The single PubMedBERT transformer scored significantly higher than the baseline system in the challenge evaluation, achieving an F1 score of 0.8493 in the strict evaluation. We extended the baseline MeSH normalization procedure, using biomedical word embeddings to try to improve recall. Our system scored slightly lower than the baseline in the strict evaluation, achieving an F1 score of 0.7870. Our entry for the automatic MeSH indexing sub-task achieved an F1 score of 0.3334.**

Keywords*— *Named Entity Recognition, Entity Normalization, Transformer, Stacking Ensemble, Indexing

I. Introduction

The NLM-Chem track of the BioCreative VII challenge is divided into two subtasks. The first subtask focuses on the automatic identification of chemical entities in biomedical literature and their subsequent normalization to Medical Subject Heading (MeSH) codes. The second subtask is related to automatic document level MeSH indexing (1). The NLM-Chem dataset (2, 3) is made up of 150 full PubMed articles, with each chemical entity character span labelled, as well as any MeSH codes each entity maps to. Each article also contains a list of MeSH index terms.

In recent years transformer models have become very popular in natural language processing (NLP). Large transformer language models (e.g. BERT), trained with very large datasets, are easily repurposed for novel tasks via transfer learning, with impressive results (4, 5). The adaptation of general domain transformer models to the biomedical domain is an active research area. Initial efforts involved continuing the BERT pre-training process using biomedical corpora (6, 7). Gu et al. went a step further by training a BERT model from scratch using biomedical corpora (PubMedBERT), and demonstrated this approach improved performance across a range of tasks (8). We chose to work with this model (PubMedBERT) for the chemical entity recognition challenge component. In addition to the NLM-Chem dataset, we made use of two chemical NER datasets from previous BioCreative challenges: the CDR, and CHEMDNER datasets (9, 10). The three datasets have distinct annotation guidelines and therefore can't easily be combined. We aimed to investigate whether a secondary model, trained using the outputs of separate models, each fine-tuned using one of the three datasets, could improve on the performance of a single model, trained using the NLM-Chem dataset only.

The paper introducing the NLM-Chem dataset (2) presents baseline methods for chemical entity recognition and subsequent normalization to MeSH IDs. Their normalization method is based on the sieve approach (11). In this work, we aimed to improve the baseline normalization approach's recall by identifying predicted entity synonyms using a biomedical word embedding model (12).

For the indexing task, we trained a logistic regression model to predict whether a given MeSH code – identified using the automatic chemical NER and normalization systems - appears in the article index based on its mention frequency, identity, and the document metadata.

II. Method

We fine-tuned separate PubMedBERT models using the NLM-Chem, CDR, and CHEMDNER datasets. Table 1 lists the model and training parameters. A subset of the parameters were tuned using the validation sets included with each of the datasets - the token level micro-averaged F1 score was used to select the optimal parameters. Following parameter tuning, each model was fine-tuned again using the optimal parameters with early stopping. When training with early stopping, each model was evaluated using the validation set at intervals of 100 iterations, and training was terminated when the token level micro-averaged F1 score failed to improve after 500 iterations. The best performing model checkpoint was selected for the evaluation. Each model was trained for a maximum of 10 epochs, but training terminated before then during each training run. Note that PubMedBERT is able to process sequences containing a maximum of 512 tokens. All training and validation sequences were truncated to this length prior to model training. Next, we generated predictions for the NLM-Chem dataset using each of the models. We then trained several neural network stacking models to combine the individual model predictions at the token level using the NLM-Chem dataset. We trained the stacking models with a

TABLE 1: MODEL/TRAINING PARAMETERS. PAREMETERS WITH MULTIPLE ENTRIES WERE TUNED DURING OUR EXPERIMENTS.

PubMedBERT	
Batch size	16, 32
Epochs	2, 3, 4, 5
Learning rate	1e-5, 3e-5, 5e-5
Learning rate schedule	Linear, warm-up ratio = 0.1
Optimizer	AdamW (beta1=0.9, beta2=0.999, epsilon=1e-8 decay=0.01)
Max gradient norm	1.0
Stacking Model	
Dropout	0.0, 0.01, 0.025, 0.05, 0.10
L2 coefficient	0.0, 0.01, 0.05
# ReLU layers	0, 1
# Hidden units	0, 1024
Optimizer	Adam (beta1=0.9, beta2=0.999)
Learning rate	0.001
Loss	Categorical crossentropy

batch size of 1024 for a maximum of 500 epochs using early stopping based on validation loss, with a patience of 3 epochs. Table 1 lists the model and training parameters, some of which were tuned.

The NLM-Chem dataset contains full articles, and as a result, many of the included passages are too long to be processed by PubMedBERT. When generating predictions for the evaluation set, instead of truncating long texts, as we did when training the models, we split each text into overlapping sequences of 512 tokens, processed each of the sequences individually, and then averaged the outputs. We used a stride of 255 tokens to ensure an even coverage of backward and forward context when generating predictions (we used a stride of 255 instead of 256 because the first and final tokens are special reserved tokens).

We adopted a normalization sieve approach that is similar to those presented in (2, 11). First, we normalized each of the terms contained in the Medical Subject Headings (MeSH) database by lowercasing, removing whitespace, and substituting Greek characters with their English spellings. We then defined a set of string manipulation and substitution methods designed to improve recall when normalizing entities to MeSH codes. The string manipulation methods included stemming, and non-alphanumeric character removal. The substitution methods included abbreviation expansion using Ab3p (13) and similar entity substitution using BioWordVec embeddings (12). Abbreviation expansion involves identifying abbreviations and resolving them to their long form, which is typically less ambiguous. The similar entity substitution method using BioWordVec embeddings works as follows: given an entity, the ten most similar BioWordVec strings – according to cosine similarity - are retrieved, strings with a similarity score below some threshold are discarded (the optimal threshold was determined empirically), the remaining strings are then resolved to MeSH codes where possible, and the final MeSH code is determined by majority vote. We made use of several synonym data sources in addition to the MeSH database: ChEBI (14), UMLS (15), NCI Thesaurus (16), UNII (https://fdasis.nlm.nih.gov/), and PubChem (17). We evaluated the normalization precision of each of the data source and string manipulation/substitution method combinations and used this information to construct the normalization sieve.

For the indexing subtask we trained a logistic regression model to predict whether an identified MeSH code belongs in the article's index. The model used the following features: mention count, normalized mention count, number of different passages the code is mentioned in, passage type mention count (e.g. code is mentioned 3 times in the abstract), and the code identity.

III. RESULTS

Table 2 compares the chemical entity span identification performance of the single PubMedBERT model with the stacking model (trained using outputs from the NLMChem, CDR, and CHEMDNER models). The models performed equally well, therefore we chose the simpler, single PubMedBERT model when generating predictions for the challenge submission. Note that before generating the final test set predictions, we re-trained the model using the optimal configuration and early stopping with a new training set combining the original training and test sets. Our system's performance is compared with the baseline system in Table 3 and Table 4. Our chemical NER model significantly outperformed the baseline: its recall was considerably higher while its precision was only slightly lower. Our system performed slightly worse than the baseline in the strict evaluation and scored significantly worse in the approximate evaluation. Our system scored higher in recall, with a lower precision in both evaluations when compared with the baseline. It's worth noting that our normalization procedure was evaluated using the output of a more effective entity recognition system, therefore the evaluation is biased in favour of our system, and the baseline normalization procedure may outperform ours by a greater margin when normalizing the same input.

Table 5 compares the normalization sieve performance on the NLM-Chem dataset with different components removed both for the full ground truth dataset, and the model's validation set predictions. Including additional data sources produced a large performance improvement: recall was improved dramatically with a comparatively small reduction in precision. String manipulation is the next most important component – removing this harms F1 score in both the ground truth dataset and the validation set predictions. The impact of removing the abbreviation expansion and similar entity components is less clear cut: the difference in F1 score with and without these components is very small. Even so, we included both components in the normalization procedure because we observed no evidence that they harmed normalization performance.

Table 6 compares our automatic indexing system with the baseline system. Our system is more precise but with a much lower recall, resulting in a lower F1 score. Neither system performs well, reflecting the difficulty of the task. Two major shortcomings of our approach are: it doesn't exploit the article content, and it assumes the index terms are independent. Future work could address these issues.

IV. Conclusion

We compared two different approaches to chemical NER: a single PubMedBERT model fine-tuned using the NLM-Chem dataset, and a stacking model combining three separate PubMedBERT models, each fine-tuned with one of three datasets: NLM-Chem, CHEMDNER, and CDR. The stacking model didn't improve upon the single PubMedBERT model, so we chose a single fine-tuned PubMedBERT model for the final evaluation. The single PubMedBERT model significantly outperformed the baseline. We extended the baseline normalization sieve approach with an additional component designed to improve recall by identifying similar chemical entities using the BioWordVec biomedical word embedding model. Our normalization approach performed slightly worse than the baseline system in the strict evaluation and scored significantly worse in the approximate evaluation. We trained a logistic regression model to identify index MeSH codes. It is more precise than the baseline system, but achieved a significantly lower recall, and a lower F1 score.

TABLE 2: COMPARISON OF FINE-TUNED PUBMEDBERT MODEL AND STACKING MODEL, EVALUATED USING THE NLM-CHEM TEST SET.

Model	**Strict evaluation**		
	Precision	***Recall***	***F1***
PubMedBERT	0.806	0.854	0.829
Stacking model	0.795	0.860	0.827

TABLE 3: CHALLENGE EVALUATION RESULTS - STRICT/APPROXIMATE CHEMICAL ENTITY SPAN IDENTIFICATION.

System	**Chemical Entity Span Identification**					
	Strict evaluation			**Approximate evaluation**		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
Baseline	0.8440	0.7877	0.8149	0.9156	0.8492	0.8811
Ours	0.8338	0.8654	0.8493	0.8953	0.9309	0.9127

TABLE 4: CHALLENGE EVALUATION RESULTS – STRICT/APPROXIMATE ENTITY NORMALIZATION.

System	**Chemical Entity Normalization**					
	Strict evaluation			**Approximate evaluation**		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
Baseline	0.8151	0.7644	0.7889	0.7917	0.7889	0.7857
Ours	0.7890	0.7849	0.7870	0.7192	0.8254	0.7628

TABLE 5: NORMALIZATION SIEVE ABLATION STUDY.

Ablation details	**NLM-Chem ground truth**			**Validation set predictions**		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
MeSH data only	0.915	0.682	0.781	0.839	0.667	0.743
No string manipulation	0.880	0.801	0.839	0.808	0.779	0.793
No abbreviation expansion	0.877	0.808	0.841	0.805	0.790	0.797
No similar entity matching	0.885	0.801	0.841	0.815	0.787	0.80
Complete sieve	0.881	0.808	0.843	0.809	0.790	0.799

TABLE 6: CHALLENGE EVALUATION RESULTS – STRICT/APPROXIMATE INDEXING.

System	Chemical Entity Normalization					
	Strict evaluation			Approximate evaluation		
	Precision	***Recall***	***F1***	***Precision***	***Recall***	***F1***
Baseline	0.3134	0.6101	0.4141	0.4510	0.7816	0.5329
Ours	0.4073	0.2822	0.3334	0.4844	0.4612	0.4380

REFERENCES

1. Leaman,R., Islamaj,R. and Lu,Z. 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. 2021.

2. Islamaj,R., Leaman,R., Kim,S., Kwon,D., Wei,C.H., Comeau,D.C., Peng,Y., Cissel,D., Coss,C., Fisher,C., Guzman,R., Gokal Kochar,P., Koppel,S., Trinh,D., Sekiya,K., Ward,J., Whitman,D., Schmidt,S. and Lu,Z. (2021) NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci Data* **8**, 91.

3. Islamaj,R., Leaman,R., Cissel,D., Cheng,M., Coss,C., Denicola,J., Fisher,C., Guzman,R., Kochar,P., Miliaras,N., Punske,Z., Sekiya,K., Trinh,D., Whitman,D., Schmidt,S. and Lu,Z. 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. 2021.

4. Devlin,J., Chang,M.W., Lee,K. and Toutanova,K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

5. Wolf,T., Debut,L., Sanh,V., Chaumond,J., Delangue,C., Moi,A., Cistac,P., Rault,T., Louf,R., Funtowicz,M., Davison,J., Shleifer,S., von Platen,P., Ma,C., Jernite,Y., Plu,J., Xu,C., Le Scao,T., Gugger,S., Drame,M., Lhoest,Q. and Rush,A.M. (2020) HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*.

6. Peng,Y., Yan,S. and Lu,Z. (2019) Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv:1906.05474*.

7. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., Ho So,C., and Kang,J. (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv:1901.08746*.

8. Gu,Y., Tinn,R., Cheng,H., Lucas,M., Usuyama,N., Liu,X., Naumann,T., Gao,J. and Poon,H. (2020) Domain-Specific Language Model Pretraining for Biomedical Natural Language. *arXiv:2007.15779*.

9. Li,J., Sun,Y., Johnson,R.J., Sciaky,D., Wei,C.H., Leaman,R., Peter Davis,A., Mattingly,C.J., Wiegers,T.C. and Lu,Z. (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* (Oxford). 2016:baw068.

10. Krallinger,M., Rabal,O., Leitner,F., Vazquez,M., Salgado,D., Lu,Z., Leaman,R., Lu,Y., Ji,D., Lowe,D.M., Sayle,R.A., Batista-Navarro,R.T., Rak,R., Huber,T., Rocktäschel,T., Matos,S., Campos,D., Tang,B., Xu,H., Munkhdalai,T., Ryu,K.H., Ramanan,S.V., Nathan,S., Žitnik,S., Bajec,M., Weber,L., Irmer,M., Akhondi,S.A., Kors,J.A., Xu,S., An,X., Sikdar,U.K., Ekbal,A., Yoshioka,M., Dieb,T.M., Choi,M., Verspoor,K., Khabsa,M., Giles,C.L., Liu,H., Ravikumar,K.E., Lamurias,A., Couto,F.M, Dai,H.J., Tsai,R.T.H., Ata,C., Can,T., Usié,A., Alves,R., Segura-Bedmar,I., Martínez,P., Oyarzabal,J. and Valencia,A. (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J Cheminform* **7**, S2.

11. D'Souza,J. and Ng,V. (2015) Sieve-Based Entity Linking for the Biomedical Domain. *ACL*.

12. Zhang,Y., Chen,Q., Yang,Z., Lin,H. and Lu,Z. (2019) BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data* **6**, 52.

13. Sohn,S., Comeau,D.C., Kim,W. and Wilbur,W.J. (2008) Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, **9**, 402.

14. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G. Turner,S., Williams,M. and Steinbeck,C. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* **41**, D456–463.

15. Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32**, D267–270.

16. Sioutos,N., de Coronado,S., Haber,M.W., Hartel,F.W., Shaiu,W.L. and Wright,L.W. (2007) NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics* **40**, 30-43.

17. PubChem [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2004 - [cited 2021 Sep 20]. Available from: https://www.ncbi.nlm.nih.gov/geo/