AIONER: All-in-one scheme-based biomedical named entity recognition using deep learning

(Supplementary Information)

Table S1. The existing datasets within the six popular concepts in biomedical literature.

Dataset	Gene	Disease	Chemical	Variant	Species	CellLine
BioRED (Luo *et al.*, 2022)	○	○	○	○	○	○
CRAFT (Bada *et al.*, 2012)	○		○			○
GNormPlus (Wei *et al.*, 2015)	○					
NLM-Gene (Islamaj *et al.*, 2021)	○					
BioCreative II GM (Smith *et al.*, 2008)	○					
ChemProt (Krallinger *et al.*, 2017)	○		○			
DrugProt (Miranda *et al.*, 2021)	○		○			
JNLPBA (Kim *et al.*, 2004)	○					○
GPRO (Krallinger *et al.*, 2015)	○					
RENET (Wu *et al.*, 2019)	○	○				
RENET2 (Su *et al.*, 2021)	○	○				
NCBI Disease (Doğan *et al.*, 2014)		○				
BC5CDR (Li *et al.*, 2016)		○	○			
NLM-Chem (Islamaj *et al.*, 2021)			○			
CHEMDNER (Krallinger *et al.*, 2015)			○			
CEMP (Krallinger *et al.*, 2015)			○			
tmVar1-3 (Wei *et al.*, 2022; Wei *et al.*, 2013; Wei *et al.*, 2018)				○		
Nala (Cejuela *et al.*, 2017)				○		
SETH (Thomas *et al.*, 2016)				○		
OSIRIS (Furlong *et al.*, 2008)				○		
Linnaeus (Gerner *et al.*, 2010)					○	
Species-800 (Pafilis *et al.*, 2013)					○	
BioID corpus (Arighi *et al.*, 2017)						○
CellFinder (Kaewphan *et al.*, 2016)						○

Table S2. The corpora for the evaluation of BioNER tasks with new entity types.

Task	# Entity type	Text genre	Text size	# Entity
DMCB_Plant (Cho *et al.*, 2017)	1	PubMed abstract	208	3,985
AnEM (Pyysalo and Ananiadou, 2014)	11	PubMed abstract + PMC full-text	500	3,135
BEAR (Wührl and Klinger, 2022)	14	Twitter	2,100	6,324

Table S3. Comparison of different task-specific tagging modes on the BioRED test set. We examine the task-specific tagging modes for identifying all entities in the BioRED test set. The all-in-one (AIO) mode employs the "<ALL></ALL>" to directly recognize all concept entities. On the other hand, the individual (IND) mode predicts the corresponding entity type by using each individual tag (e.g., <Gene></Gene> for gene type) and then unified the results to compute the performance. When the predicted entities overlap, we select the longest entity as the final result. To investigate if the performance of the IND and AIO modes can be improved by combining their results, we integrated the output of both modes and included the combined results in the last row of the table (denoted as "Combined"). The table presented below illustrates the results, indicating that the AIO mode achieves higher F1-scores for overall performance and each entity type on BioRED. This is primarily because the entity scope and definition in individual corpora do not align completely with BioRED. For instance, Linnaeus and Species-800 do not annotate species relevant clinical terms, such as "patient". Combining all predicted results of IND and AIO did not improve the performance, leading us to recommend using the AIO tag mode to identify all concept entities in BioRED and the IND marks to recognize the entities in the individual corpora.

Task-specific tagging mode	Overall	Gene	Disease	Chemical	Species	Variant	CellLine
IND mode	86.37	91.26	87.27	88.08	57.24	88.02	88.17
AIO mode	**91.26**	**92.40**	**88.07**	**90.98**	**97.50**	**88.51**	**90.53**
Combined	90.42	91.73	87.28	89.48	97.50	87.37	90.32

Experimental Setup and Hyper-parameter Tuning

In terms of hyperparameters, we focused on optimizing the learning rate, and did not perform any additional optimization. Specifically, we randomly selected 10% of the training set as a development set to determine the best learning rate, selecting from a set of four possible options: 1e-5, 5e-5, 5e-6, and 1e-6. The learning rate was chosen based on the highest F1-score on the development set, and the development set was then combined with the training set for subsequent model evaluation. The number of training epochs was determined using the patience parameter, which was set to 5, with a maximum of 50 epochs. The training was stopped if there was no improvement in accuracy after five consecutive epochs. To ensure fairness in comparison, each developed model was optimized using this method.

The primary hyperparameters of AIONER were set as follows: a learning rate of 5e-6, batch size of 32, and a maximum input length of 256 tokens. Our NER models take sentences as input, and based on the training data we used, only 0.04% of sentences were longer than 256 tokens. Setting the maximum input length to 256 tokens allows the model to effectively cover most sentence lengths, and helps to ensure more efficient training and testing. We also tested a maximum input length of 512 tokens, but found that PubMedBERT did not perform better than the model with a length of 256 tokens (91.01% vs. 91.26%).

Reference

Arighi, C. *et al.* (2017) Bio-ID track overview. In *BioCreative VI Workshop*, 28–31.

Bada, M. *et al.* (2012) Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, **13**, 1-20.

Cejuela, J.M. *et al.* (2017) nala: text mining natural language mutation mentions. *Bioinformatics*, **33**, 1852-1858.

Cho, H. *et al.* (2017) A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*, **18**, 1-12.

Doğan, R.I. *et al.* (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. *J. Biomed. Inform.*, **47**, 1-10.

Furlong, L.I. *et al.* (2008) OSIRISv1. 2: a named entity recognition system for sequence variants of genes in biomedical literature. *BMC bioinformatics*, **9**, 1-16.

Gerner, M. *et al.* (2010) LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, **11**, 1-17.

Islamaj, R. *et al.* (2021) NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, **8**, 1-12.

Islamaj, R. *et al.* (2021) NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *J. Biomed. Inform.*, **118**, 103779.

Kaewphan, S. *et al.* (2016) Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, **32**, 276-282.

Kim, J.-D. *et al.* (2004) Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, 70-75.

Krallinger, M. *et al.* (2017) Overview of the BioCreative VI chemical-protein interaction Track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, 141-146.

Krallinger, M. *et al.* (2015) Overview of the CHEMDNER patents task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, 63-75.

Li, J. *et al.* (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, **2016**, baw068.

Luo, L. *et al.* (2022) BioRED: a rich biomedical relation extraction dataset. *Brief. Bioinform.*, **23**, bbac282.

Miranda, A. *et al.* (2021) Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of druggene/protein relations *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Pafilis, E. *et al.* (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One*, **8**, e65390.

Pyysalo, S. and Ananiadou, S. (2014) Anatomical entity mention recognition at literature scale. *Bioinformatics*, **30**, 868-875.

Smith, L. *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome biology*, **9**, 1-19.

Su, J. *et al.* (2021) RENET2: high-performance full-text gene–disease relation extraction with iterative training data expansion. *NAR Genomics and Bioinformatics*, **3**, lqab062.

Thomas, P. *et al.* (2016) SETH detects and normalizes genetic variants in text. *Bioinformatics*, **32**, 2883-2885.

Wei, C.-H. *et al.* (2022) tmVar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics*, btac537.

Wei, C.-H. *et al.* (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, **29**, 1433-1439.

Wei, C.-H. *et al.* (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, **2015**, 918710

Wei, C.-H. *et al.* (2018) tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*, **34**, 80-87.

Wu, Y. *et al.* (2019) Renet: A deep learning approach for extracting gene-disease associations from literature. In *International Conference on Research in Computational Molecular Biology*, 272-284.

Wührl, A. and Klinger, R. (2022) Recovering Patient Journeys: A Corpus of Biomedical Entities and Relations on Twitter (BEAR). In *Proceedings of the Language Resources and Evaluation Conference*, 4439-4450.