Exploring Biomedical Relation Extraction through ChatGPT Augmentation and Dual Training

Han-Ting Yu[1], Bo-Cheng Qiu[2], Shao-Ting Yen[3], Cheng-Yang Wang[1], Yu-Han Wu[4], Shao-Man Lee[1*], and Yi-Yu Hsu[1*]

[1] Miin Wu School of Computing, National Cheng Kung University, Tainan, Taiwan
[2] Dept. of Statistics, National Cheng Kung University, Tainan, Taiwan
[3] Cross College Elite Program, National Cheng Kung University, Tainan, Taiwan
[4] Dept. of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

*Corresponding authors: E-mail: shaomanlee@gs.ncku.edu.tw and yiyuhsu@gs.ncku.edu.tw

Abstract

Relation extraction in biomedical text mining faces challenges due to complex terminology and rapidly growing literature. Our research focuses on improving relation extraction through data augmentation and targeted dual training. We fine-tuned a PubMedBERT model and enriched it with GPT-4 generated examples, iteratively refining the process. A dual training approach focusing on chemical entities significantly improved F1 scores by 2.04% on average. Our strategies demonstrate the effectiveness of ChatGPT-based augmentation and selective dual training for advancing biomedical text mining.

Introduction

The discipline of biomedical text mining confronts a myriad of challenges, intricately woven into the specialized and often arcane nature of biomedical literature. These challenges encompass specialized vocabularies, multifaceted terminological frameworks, and the unremitting expansion of scholarly contributions. A multitude of computational algorithms and models—such as PubMedBERT (1), BioBERT (2), AIONER (3), and BioSyn (4)—have been marshaled to mitigate these issues, with particular focus on Named Entity Recognition (NER) and Named Entity Normalization (NEN).

Our research furthers the field by pioneering novel tokenization methodologies and harmonizing specialized NER models, specifically BioBERT and AIONER, to amplify NER accuracy. Additionally, this research accentuates the domain of Relation Extraction (RE), wherein we implement data augmentation paradigms and harness the computational prowess of GPT-4. Empirical evidence substantiates the efficaciousness of these strategies, engendering an enhancement in RE performance metrics. Notably, a nuanced dual training schema has been implemented, which significantly amplifies predictive accuracy for chemical entities, thereby optimizing the overall efficiency of RE processes.

Methods

Our methodology utilizes the BioRED repository to enhance systems for biomedical relation extraction across three stages:

Named Entity Recognition (NER)

In the NER phase, we identified specific anomalies in tokenization practices, exemplified by the erroneous separation of terms like "p.G380R" into "p." and "G380R". To address this issue, we deployed nltk.tokenize.RegexTokenizer, which was adapted with customized rule sets for meticulous tokenization. Given the prevalence of specialized biomedical terminology in the dataset, we incorporated BioBERT, tailored explicitly for biomedical literature, along with AIONER, which was rigorously trained on comparable data.

Named Entity Normalization (NEN)

During the NEN stage, unique identifiers are ascribed to entities previously identified by NER. We leverage the BioSyn model, designed to optimize synonymic representations among top candidate lexemes. This model is synergistically combined with a curated lexicon, and a table comprising entity identifiers sourced from both the PubTator API and the BioCreative dataset is created. Upon identification of an entity by NER, its identifier is predicted through BioSyn and cross-referenced with our curated table, effecting replacements as warranted.

Relation Extraction (RE)

For the RE component, the PubMedBERT model was fine-tuned utilizing a dataset from the BioCreative competition, which comprised 400 documents for training, 100 for development, and a distinct validation set of 1,000 documents. To augment this dataset, GPT-4 was enlisted, and a streamlined workflow was established to ensure data coherence. Figure 1 delineates the intricate process involved in GPT-4-augmented biomedical relationship extraction.

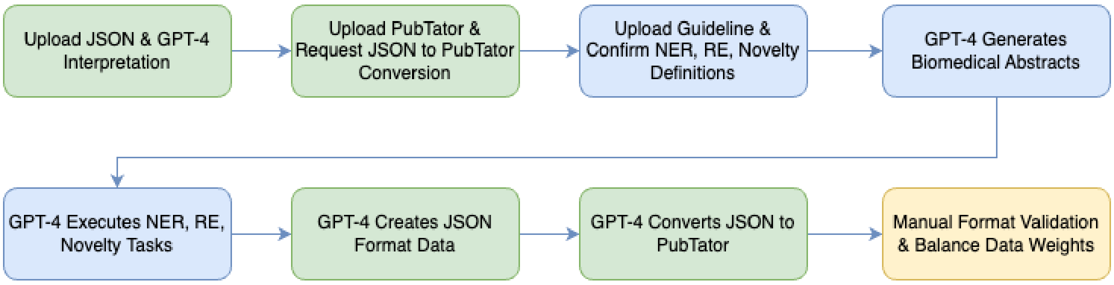
Upload JSON & GPT-4 Interpretation
Upload PubTator & Request JSON to PubTator Conversion
Upload Guideline & Confirm NER, RE, Novelty Definitions
GPT-4 Generates Biomedical Abstracts
GPT-4 Executes NER, RE, Novelty Tasks
GPT-4 Creates JSON Format Data
GPT-4 Converts JSON to PubTator
Manual Format Validation & Balance Data Weights

Figure 1. A Schematic Outline of GPT-4-Augmented Biomedical Relationship Extraction: The process involves data input, integration of PubTator data, guideline establishment enriched by GPT-4, AI-generated abstracts, NLP system execution, data structuring, format transformation with GPT-4 enhancements, and quality assurance with GPT-4 reinforcement.

In the initial data preparation stage, we used GPT-4 to manage JSON files and adopted a segmented strategy for better efficiency. Guidelines from the "BioRED_Annotation_Guideline.pdf" informed our protocols for Named Entity Recognition (NER) and Relation Extraction (RE).

GPT-4 was calibrated to align with our research goals using predefined queries and prompts. Feedback from this process led to minor adjustments in the training dataset, improving the model's performance.

The research thus addresses and provides practical solutions to prevalent challenges, including data offset inaccuracies and formatting inconsistencies. Data augmentation and generative content creation were utilized to enrich the dataset. Moreover, a balanced label distribution, set at a 4:1 ratio between training and validation sets, contributed to enhanced model performance.

Results

We experimented with different augmentation and balancing approaches across five runs of relation extraction.

(1) A balanced dataset without BioRED dataset augmentation.
(2) An unbalanced dataset with BioRED dataset augmentation.
(3) A balanced dataset with BioRED dataset augmentation.
(4) A balanced dataset with BioRED dataset augmentation and removing 100 documents from the gold standard.
(5) A balanced dataset with BioRED dataset augmentation, enriched by incorporating data from GPT-4.

Model 2 performed well on "Relation Type" and "Novelty." Model 5 with GPT-4 augmentation achieved the highest precision for "Entity Pair + Relation Type + Novelty." (Table 1).

Table 1. RE model performance of official runs. *APR: Average of participants runs; *MPR: Median of participants runs

Model	Entity Pair (%)			+Relation Type (%)			+Novelty (%)		
	P	R	F	P	R	F	P	R	F
1	73.93	71.48	72.68	49.82	48.17	48.98	36.49	35.28	35.88
2	75.63	72.49	74.03	54.46	52.20	53.31	41.61	39.88	40.73
3	73.45	74.73	74.08	52.43	53.35	52.88	40.14	40.84	40.49
4	73.43	74.75	74.08	52.42	53.36	52.89	40.12	40.84	40.48
5	73.32	74.86	74.08	52.28	53.38	52.82	40.13	40.98	40.55
APR	69.22	68.60	67.03	49.01	48.39	47.74	36.15	35.73	35.22
MPR	77.93	69.65	73.56	51.64	54.79	53.17	41.61	39.88	40.73

Combining NER and RE models yielded additional runs (Table 2):

(1) dmis-lab/biobert-large-cased-v1.1-mnli with RE Model 3.
(2) dmis-lab/biobert-large-cased-v1.1-mnli with RE Model 1.
(3) alvaroalon2/biobert_diseases_ner with RE Model 3.
(4) alvaroalon2/biobert_diseases_ner with RE Model 1.
(5) AIONER with RE Model 3.

Table 2. NER and RE model of official runs. *APR: Average of participants runs; *MPR: Median of participants runs

Model	NER (%)			ID (%)			Entity Pair (%)			+Relation Type (%)			+Novelty (%)		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	78.29	55.74	65.12	53.75	51.27	52.48	21.26	13.55	16.55	16.18	10.41	12.67	12.04	7.79	9.46
2	78.29	55.74	65.12	53.75	51.27	52.48	21.43	12.95	16.14	15.48	9.53	11.80	11.18	6.91	5.54
3	72.95	53.95	62.03	62.03	47.73	49.28	18.17	12.78	15.01	13.84	9.88	11.53	10.47	7.50	8.74
4	72.95	53.95	62.03	62.03	47.73	49.28	18.86	12.30	14.89	13.76	9.20	11.03	9.92	6.66	7.97
5	86.43	88.51	87.46	46.80	55.76	50.89	12.81	16.02	14.24	9.61	12.09	10.71	7.26	9.15	8.10
APR	80.38	74.48	76.87	65.57	62.67	63.36	34.14	26.48	28.62	25.13	19.87	21.39	19.00	15.12	16.25
MPR	83.35	74.33	78.58	69.02	64.75	66.81	30.14	40.26	34.47	22.15	29.75	25.40	17.18	23.33	19.79

The competition model results in Tables 1 and 2 reveal areas needing improvement. Model 1 requires enhancement across all metrics, especially for the augmented dataset. Model 2 achieved the highest precision of 41.61% on "Entity Pair + Relation Type + Novelty", performing well on "Relation Type" and "Novelty." However, Models 3-5 exhibited limitations in precision without sacrificing recall. Further analysis of Model 2's effective strategies could inform refinements to improve precision. The training set balancing approach may also require adjustment based on Model 2's performance.

To enhance predictive accuracy, we implemented a dual training strategy focused on relation pairs containing chemical entities. In this approach, chemical entity pairs were trained twice, while other relation pairs were trained once. This selective dual training targeted chemical entities due to their complexity arising from specialized naming conventions and terminology.

Performance metrics for the unofficial runs utilizing dual training are detailed in Tables 3 and 4. Compared to the single training approach, dual training led to notable improvements, including a 1.14 point increase in F1 for "Entity Pair", a 2.55 point enhancement for "Entity Pair + Relation Type", and a 2.42 point gain for "Entity Pair + Relation Type + Novelty." By selectively doubling the instances of complex chemical entities, dual training significantly boosted performance on key metrics.

Table 3: RE model performance of unofficial runs.

Model	Entity Pair (%)			+Relation Type (%)			+Novelty (%)		
	P	R	F	P	R	F	P	R	F
1	73.29	74.86	74.07	52.26	53.38	52.81	40.12	40.98	40.54
2	75.78	73.37	74.56	54.26	52.53	53.38	42.05	40.71	41.37
3	74.77	75.56	75.16	55.54	56.13	55.84	43.14	43.60	43.37
4	76.80	72.32	74.49	57.49	54.14	55.77	43.53	40.99	42.22
5	76.21	74.58	75.39	56.45	55.24	55.84	43.20	42.27	42.73

Table 4: NER and RE model performance of unofficial runs.

Model	NER (%)			ID (%)			Entity Pair (%)			+Relation Type (%)			+Novelty (%)		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1	86.43	88.51	87.46	46.80	55.76	50.89	13.22	15.54	14.28	9.93	11.77	10.77	7.63	9.07	8.29
2	86.43	88.51	87.46	46.80	55.76	50.89	12.64	16.15	14.18	9.81	12.62	11.04	7.53	9.71	8.48
3	86.43	88.51	87.46	46.80	55.76	50.89	13.33	15.14	14.18	10.19	11.64	10.87	7.55	8.65	8.06
4	86.43	88.51	87.46	46.80	55.76	50.89	13.12	15.57	14.24	10.08	12.04	10.97	7.59	9.10	8.28

Conclusion

Our strategies demonstrate the value of ChatGPT augmentation and targeted dual training for chemical entities, advancing biomedical relation extraction. Future work should explore balancing strategies inspired by Model 2's performance.

Funding

This work was supported by NSTC 112-2321-B-006-008.

References

1. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

2. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

3. Luo, L., Wei, C. H., Lai, P. T., Leaman, R., Chen, Q., and Lu, Z. (2023). AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. Bioinformatics, 39(5), btad310.

4. Sung, M., Jeon, H., Lee, J., and Kang, J. (2020). Biomedical entity representations with synonym marginalization. arXiv preprint arXiv:2005.00239.

5. Luo, L., Lai, P. T., Wei, C. H., Arighi, C. N., and Lu, Z. (2022). BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics, 23(5), bbac282.