Puer at SemEval-2024 Task2: A BioLinkBERT Approach to Biomedical Natural Language Inference

Jiaxu Dao, Zhuoying Li, Xiuzhong Tang, Xiaoli Lan, Junde Wang
School of Technology
Pu'er University
{daojiaxu, lizhuoying, tangxiuzhong, lanxiaoli, wangjunde}@peu.edu.cn

Abstract

This paper delineates our investigation into the application of BioLinkBERT for enhancing clinical trials, presented at SemEval-2024 Task 2. Centering on the medical biomedical NLI task, our approach utilized the BioLinkBERT-large model, refined with a pioneering mixed loss function that amalgamates contrastive learning and cross-entropy loss. This methodology demonstrably surpassed the established benchmark, securing an impressive F1 score of 0.72 and positioning our work prominently in the field. Additionally, we conducted a comparative analysis of various deep learning architectures, including BERT, ALBERT, and XLM-RoBERTa, within the context of medical text mining. The findings not only showcase our method's superior performance but also chart a course for future research in biomedical data processing. Our experiment source code is available on GitHub at: https://github.com/daojiaxu/semeval2024_task2.

1 Introduction

Clinical Trial Reports (CTRs) play a crucial role in documenting the methods and results of clinical trials(Jullien et al., 2023a; Vladika and Matthes, 2023). It contains a detailed overview of participant circumstances, intervention experiment descriptions, experimental results, and adverse events that happened in the participants. Natural Language Inference is a valuable technique for analyzing experimental data in CTR and interpreting the results. Natural Language Inference is able to analyze logical linkages, consistency, and contradictions in a document. It can assist detect logical relationships in text automatically, identify potential conflict areas fast, and improve decision-making accuracy and efficiency. Researchers can better gather and analyze clinical trial data by using Natural Language Inference techniques, which promotes medical quality improvement(Jullien et al., 2023b).

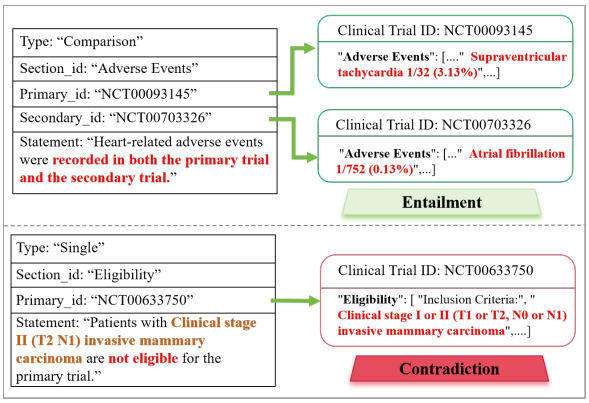

Figure 1: Dataset Example

The Figure 1 shows the example dataset used in this work. The dataset includes two forms of CTR: single and comparison. A single type CTR can retrieve relevant evidence using a Primary Id. To retrieve two relevant pieces of evidence using comparison type CTR, Primary Id and Secondary Id must be used simultaneously.

As an illustration, in the first instance, CTR represents "Heart-related adverse events were recorded in both the primary trial and the secondary trial." Searching for the matching components of the two pieces of evidence reveals that there are heart-related adverse effects, such as supraventricular tachycardia and atrial fibrosis. As a consequence, the first example is labeled as "Entailment"(Alsuhaibani, 2023). In a comparable way, in the second example, CTR believes that "Patients with clinical stage II (T2 N1) invasive breast cancer are not eligible for the primary trial." However, the participation conditions in the gathered evidence clearly show that individuals with clinical stage I or II (T1 or T2, N0 or N1) invasive mammary carcinoma match the criteria. As a result, the second case is labeled "Contradiction"(Liu et al., 2021; Zhou et al., 2023).

In the quest to push the frontiers of biomedical natural language understanding, SemEval-2024

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 70–75

Task 2 has emerged as a critical arena for testing the efficacy of AI models in parsing complex medical texts(Jullien et al., 2024). Engaging with this challenge, our work utilizes BioLinkBERT to set new benchmarks in the safety and accuracy of clinical trial inference(Ida et al., 2023; Karkera et al., 2023; Kanakarajan et al., 2022). This endeavor not only underscores the significance of developing robust NLI systems but also highlights our commitment to contributing meaningful innovations to the biomedical domain(Wang et al., 2023; Mahendra et al., 2023; Pahwa and Pahwa, 2023). Through this paper, we aim to share our methodologies, findings, and the implications they hold for the broader field of medical research, hoping to inspire further advancements and collaborative efforts in this vital area of study.

We created a number of attempts using the above dataset, and the following additions were contributed to our work:

1) We have designed a new loss function by combining the ideas of cross entropy and contrastive learning. This loss function can flexibly adjust parameters according to actual situations and has strong adaptability.

2) We have performed fine-tuning on the BioLinkBERT-large model and finally ranked 15th, achieving an F1 score of 0.72, a score of 0.59 in Faithfulness, and a score of 0.64 in Consistency.

2 System Description

2.1 Data Preprocessing

For this experiment, the training dataset was segmented into four distinct categories: Statement, Section, First Evidence, and Second Evidence. To facilitate precise identification of these text segments by the BioLinkBERT-large model, we employed the token "[SEP]" as a delineator for segment segmentation. This approach ensured that the model could accurately recognize and process the varied input text paragraphs, thereby enhancing its ability to understand and interpret the context and relationships within the data. This method of data preparation was crucial in optimizing the model's performance by providing clear structural demarcations within the training set.

More precisely, we create each input sample as shown in Figure 2.

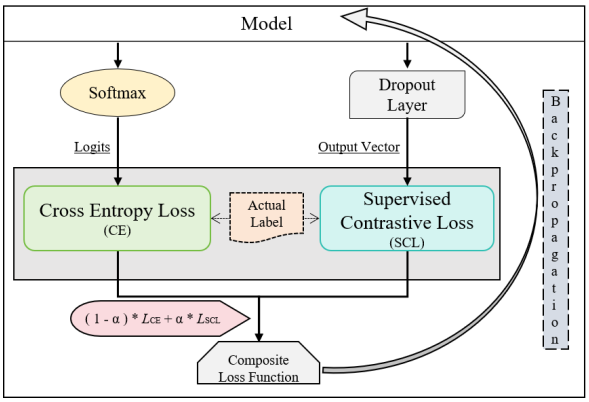

Figure 3: Composite Loss Function

Figure 2: The Architecture of Tokenizer

2.2 Model Construction

BioLinkBERT-large Model. In the domain of biological medicine, the BioLinkBERT model has been shown to be superior to the BERT model due to its ability to learn information across documents (Yasunaga et al., 2022). BioLinkBERT outperformed other models (BERT, BioMegatron, PubMedBERT, BioClinicalBERT, BioMedLM, BioGPT) in extracting the association between microorganisms and diseases from biomedical literature, with F1 precision and recall more than 0.8 (Karkera et al., 2023). The optimal accuracy was obtained in the histopathology image captioning challenge by integrating the BioLinkBERT target model with the image feature extractor ConvNexT Large (Elbedwehy et al., 2023). When compared to PubMedBERT and ChatGPT, the BioLinkBERT has demonstrated superior performance in all aspects in benchmark trials focused on biomedical text production and mining(Chen et al., 2023). The model we use is based on the BioLinkBERT large model that has been fine tuned from the MNLI and SNLI datasets.

Design of Loss Function. In the training phase, our loss function is bifurcated into two pivotal components. The initial segment utilizes the cross entropy loss function (`CrossEntropyLoss()`)(Zhang and Sabuncu, 2018), which first computes the predicted probability values via a softmax function. Subsequently, it leverages the cross entropy loss to quantify the deviation between these predicted probabilities and the actual labels, a process encapsulated by the symbol CE. The latter segment incorporates the supervised contrastive learning

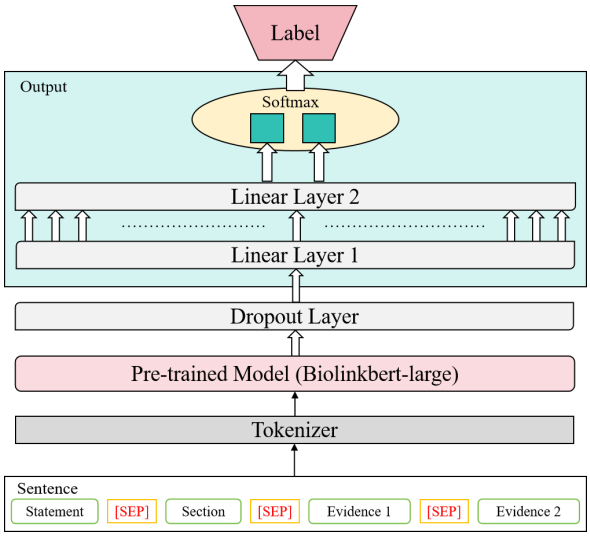

Figure 4: The Structure of System

loss function (`SupConLoss()`)(Khosla et al., 2020). Here, vectors generated post-processing by the pre-trained model are juxtaposed against the true labels to ascertain the contrastive learning loss, denoted as SCL.

Simultaneously, we have instituted a threshold parameter α to modulate the significance of each loss component. By amalgamating CE and SCL in accordance with this threshold, we obtain the composite loss. This loss is then subjected to back-propagation to minimize its magnitude, thereby aligning the predicted values more closely with the actual values. This methodology underscores our strategic approach to loss optimization, blending traditional and contrastive learning mechanisms to enhance model accuracy and performance.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log P \tag{1}$$

$$L_{SCL} = \sum_{i=1}^{N} \frac{-1}{N_{y_i} - 1} \sum_{\substack{j=1 \\ j \neq i}}^{N} 1_{y_i = y_j} \cdot \log \left(\frac{\exp(\Phi(x_i) \cdot \Phi(x_j)/\tau)}{\sum_{\substack{k=1 \\ k \neq i}}^{N} \exp(\Phi(x_i) \cdot \Phi(x_k)/\tau)} \right) \tag{2}$$

$$Loss = (1 - \alpha) * L_{CE} + \alpha * L_{SCL} \tag{3}$$

The Supervised Contrastive Learning (SCL) loss, as delineated in Equation (2), plays a pivotal role in the model's learning process by promoting the aggregation of examples from the same class while concurrently driving apart examples from distinct classes. Within a given batch, examples are meticulously grouped based on their corresponding labels, ensuring that the learning process is finely attuned to the nuances of class similarity and diversity. This is achieved through the implementation of the indicator function $1_{y_i=y_j}$, which is designed to ensure that the loss calculation exclusively considers pairs of examples (i, j) that, while sharing the same label, are distinct entities $(i \neq j)$. This deliberate focus on fostering intra-class cohesion and inter-class distinction is fundamental to augmenting the model's discriminative capabilities. A critical aspect of this approach is the use of N_{y_i}, which denotes the count of examples within the batch that share the same label as example i. This count is instrumental in normalizing the contribution of positive pairs to the loss, thereby ensuring that the SCL loss effectively enhances the model's proficiency in distinguishing between classes. This proficiency is further reinforced by the SCL loss's capacity to adjust based on the relative distances of examples within the embedding space, taking into account both positive pairs (belonging to the same class) and negative pairs (belonging to different classes), with N_{y_i} playing a crucial role in normalizing these effects based on the representation of each class within the batch.

This design strategy excels in leveraging annotated data to its fullest potential, significantly enhancing the model's generalization capabilities and the discriminative power of its feature representations. The cross-entropy loss function plays a pivotal role in assessing model performance by quantifying the discrepancy between predicted outputs and actual labels. Concurrently, the supervised contrastive learning loss function is instrumental in refining the discriminative capacity of feature representations, thereby bolstering classification accuracy. This dual-faceted approach not only ensures a comprehensive evaluation of model quality but also fosters a more nuanced understanding and representation of data features, which is crucial for achieving high precision in predictive tasks.

Model Layer Description. The levels in our model are as follows:

1) Sentence Input Layer: The model feeds the tokenizer with the text that was described in 2.1 as the training set.

2) Pre-trained Model Layer: To process the to-

Model	Loss	F1	Precision	Recall	Faithfulness	Consistency
bert-base-uncased	ce	0.6556	**0.956**	0.4989	0.0335	0.396
	ce+scl	0.6474	0.944	0.4926	0.0486	0.3931
albert-base	ce	0.6127	0.788	0.5012	0.1805	0.44
	ce+scl	0.6447	0.784	0.5474	0.2361	0.4951
biolinkbert-large	ce	0.7042	0.824	0.6149	0.4629	0.5971
	ce+scl	**0.7166**	0.764	**0.6749**	**0.5914**	**0.638**

Table 1: Comparative results of experiments in the test set

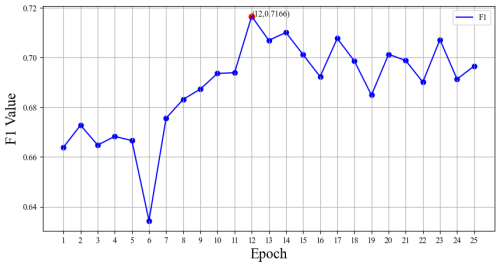

Figure 5: F1 Changes at Different Epochs on The Test Set

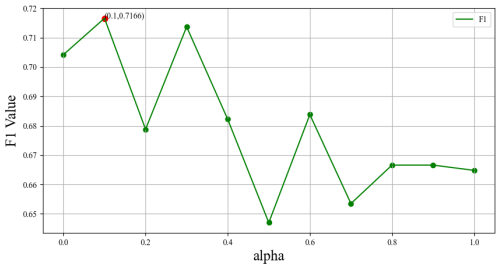

Figure 6: F1 Changes at Different Alpha on The Test Set

kenized text and produce the resultant vector representation of the text, the model makes use of the pre-trained model Biolinkbert-large.

3) Dropout layer: We implemented inactivation rate of 0.2 on the result vector to promote robustness and prevent overfitting of the model.

4) Linear Layer: To further process and transform vectors, the model employs two linear layers in the output module.

5) Softmax Function: Lastly, the model transforms the linear layer's output into a probability distribution by using the softmax function.

The loss function shown in Figure 3 was employed for backpropagation during the model training phase.The model's accuracy and real performance can be enhanced by adjusting the loss function parameter Alpha based on the current situation.

2.3 Hyper-parameters Fine-tuning

Epoch Selection. To ascertain the optimal F1 score, our experiment methodically adjusted the training duration, varying the epoch count from 1 to 20 in increments of one. At each epoch, we meticulously documented the corresponding F1 scores. As depicted by the blue line in Figure 5, a detailed analysis reveals that the F1 score peaks at epoch 12. This finding underscores the significance of epoch selection in maximizing model performance, illustrating that a carefully calibrated training period can significantly influence the effectiveness of the model's predictive accuracy.

Alpha Setting. Building upon this groundwork, we embarked on a series of experiments aimed at identifying the optimal value of alpha within the loss function, meticulously adjusting alpha from 0.1 to 1 in increments of 0.1. This systematic variation is represented by the green line in the accompanying graph. Through careful analysis, the ideal F1 score was observed when alpha was set to 0.1. This discovery not only highlights the critical role of alpha in tuning the loss function for enhanced model performance but also establishes a direct correlation between the fine-tuning of alpha and the achievement of peak predictive precision.

3 Experimental Results

In our methodology, we conducted two control trials by varying the loss function parameter Alpha, and selected three models(BERT-base-uncased(Devlin et al., 2018), ALBERT-base(Lan et al., 2019), and Biolinkbert-large)as outlined in

Table 1, aligning with the structure of our experiment. Subsequent to a rigorous examination of the experimental outcomes, it became evident that the experimental cohort employing the composite CE+SCL loss function surpassed the cohort utilizing the standalone CE loss function. This enhancement was observed across multiple metrics, including F1 score, recall, faithfulness, and consistency, specifically within the ALBERT-base and Biolinkbert-large models.

Upon comprehensive evaluation, the Biolinkbert-large model consistently demonstrates outstanding stability and superior performance. While the BERT-based-uncased model, employing the Cross-Entropy (CE) loss function, achieved the highest Precision score, it also registered relatively lower scores in terms of Faithfulness and Consistency. To encapsulate, the Biolinkbert-large model has exhibited exceptional proficiency in addressing this particular challenge.

4 Conclusion

This study has presented a comprehensive analysis of the effectiveness of BioLinkBERT in enhancing clinical trials. Our research has meticulously fine-tuned the BioLinkBERT-large model with a novel mixed loss function. The experimental results, particularly the achievement of an F1 score of 0.72, underscore the potential of leveraging advanced pre-trained language models in medical research. Our findings suggest that the integration of contrastive learning and cross-entropy loss functions significantly improves the model's performance, indicating a promising direction for future research in biomedical text mining.

Moreover, the success of this project opens new avenues for exploring the application of language models like BioLinkBERT in other domains of healthcare and medical research. Future work could focus on expanding the dataset, experimenting with different architectures, and exploring the impact of domain-specific adaptations on model performance. This could potentially lead to breakthroughs in how we process, understand, and derive insights from clinical trial reports, ultimately contributing to the advancement of medical science and patient care.

Acknowledgements

This work is supported by the 2024 Science and Technology special project of Pu'er University (PYKJZX202401 Research on medical relationship extraction task based on pre-trained language model). The authors would like to thank the anonymous reviewers for their insightful feedback.

References

Mohammed Alsuhaibani. 2023. Deep learning-based sentence embeddings using bert for textual entailment. *International Journal of Advanced Computer Science and Applications*, 14(8).

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Samar Elbedwehy, T Medhat, Taher Hamza, and Mohammed F Alrahmawy. 2023. Enhanced descriptive captioning model for histopathological patches. *Multimedia Tools and Applications*, pages 1–20.

Ryuki Ida, Makoto Miwa, and Yutaka Sasaki. 2023. Biomedical document classification with literature graph representations of bibliographies and entities. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 385–395.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023a. Nli4ct: Multi-evidence natural language inference for clinical trial reports. *arXiv preprint arXiv:2305.03598*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023b. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. *arXiv preprint arXiv:2305.02993*.

Kamal Raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. Biosimcse: Biomedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86.

Nikitha Karkera, Sathwik Acharya, and Sucheendra K Palaniappan. 2023. Leveraging pre-trained language

models for mining microbiome-disease relationships. *BMC bioinformatics*, 24(1):290.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping, et al. 2021. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592(7855):629–633.

Rahmad Mahendra, Damiano Spina, and Karin Verspoor. 2023. Ittc at semeval 2023-task 7: Document retrieval and sentence similarity for evidence retrieval in clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation, Toronto, Canada. Association for Computational Linguistics*.

Bhavish Pahwa and Bhavika Pahwa. 2023. Bphigh at semeval-2023 task 7: Can fine-tuned cross-encoders outperform gpt-3.5 in nli tasks on clinical trial data? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944.

Juraj Vladika and Florian Matthes. 2023. Sebis at semeval-2023 task 7: A joint system for natural language inference and evidence retrieval from clinical trial reports. *arXiv preprint arXiv:2304.13180*.

Weiqi Wang, Baixuan Xu, Tianqing Fang, Lirong Zhang, and Yangqiu Song. 2023. Knowcomp at semeval-2023 task 7: Fine-tuning pre-trained language models for clinical trial entailment identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1–9.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Yuxuan Zhou, Ziyu Jin, Meiwei Li, Miao Li, Xien Liu, Xinxin You, and Ji Wu. 2023. Thifly research at semeval-2023 task 7: A multi-granularity system for ctr-based textual entailment and evidence retrieval. *arXiv preprint arXiv:2306.01245*.