YNU-HPCC at SemEval-2024 Task 2: Applying DeBERTa-v3-large to Safe Biomedical Natural Language Inference for Clinical Trials

Rengui Zhang, Jin Wang and **Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, China
zrg@mail.ynu.edu.cn, {wangjin,xjzhang}@ynu.edu.cn

Abstract

This paper describes the system for the YNU-HPCC team for SemEval2024 Task 2, focusing on Safe Biomedical Natural Language Inference for Clinical Trials. The core challenge of this task lies in discerning the textual entailment relationship between Clinical Trial Reports (CTR) and statements annotated by expert annotators, including the necessity to infer the relationships in texts subjected to semantic interventions accurately. Our approach leverages a fine-tuned DeBERTa-v3-large model augmented with supervised contrastive learning and back-translation techniques. Supervised contrastive learning aims to bolster classification ac-curacy while back-translation enriches the diversity and quality of our training corpus. Our method achieves a decent F1 score. However, the results also indicate a need for further en-hancements in the system's capacity for deep semantic comprehension, highlighting areas for future refinement. The code of this paper is available at: https://github.com/RGTnuw/RG_YNU-HPCC-at-Semeval2024-Task2.

1 Introduction

Clinical trials constitute a critical component of medical research, evaluating the safety and efficacy of new treatment methods, medications, or medical devices (Avis et al., 2006). A significant num-ber of Clinical Trial Reports (CTRs) are generated throughout clinical trials. These reports typically encompass information on research design, patient demographics, treatment protocols, outcomes (such as response rates and side effects), and overall conclusions. Such comprehensive and transparent reporting of trial results provides the scientific community and the public with valuable information, informing future research and clinical practice (Zhang et al., 2020). However, the challenge is compounded by over 400,000 Clinical Trial Reports (CTRs) and their rapidly accelerating publication rate. Conducting a comprehensive review of all pertinent literature when devising treatments is impractical (DeYoung et al., 2020).

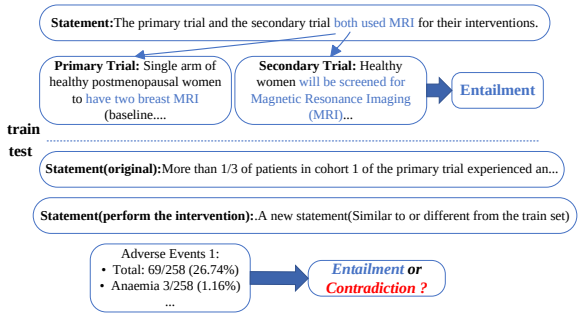

Figure 1: demonstrates textual entailment and contradiction between the medical statements and clinical trial records. Add interventions in the development and test sets.

In response to this challenge, Natural Language Inference (NLI) (Bowman et al., 2015; Devlin et al., 2019) presents a viable approach for the extensive interpretation and retrieval of medical evidence, facilitating enhanced precision and efficiency in personalized evidence-based care (Sutton et al., 2020). This task (Jullien et al., 2024) delineates the objective as classifying the inferential relationship between one or two CTR premises and a statement as either entailment or contradiction. Various interventions were applied to statements in the test and development sets, preserving or inverting entailment relations. It is imperative to ensure that inferred outcomes are justified, i.e., make correct predictions for the right reasons, and identical semantics yield consistent results, as shown in Figure 1 .

In the previous task (Jullien et al., 2023b), large language models (LLM) have achieved commendable performance(Zhou et al., 2023; Vladika and Matthes, 2023). However, the model's performance must improve when facing numerical reasoning, abbreviation, and other problems. DeBERTa-v3-large (He et al., 2023) maintained competitiveness

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 785–791

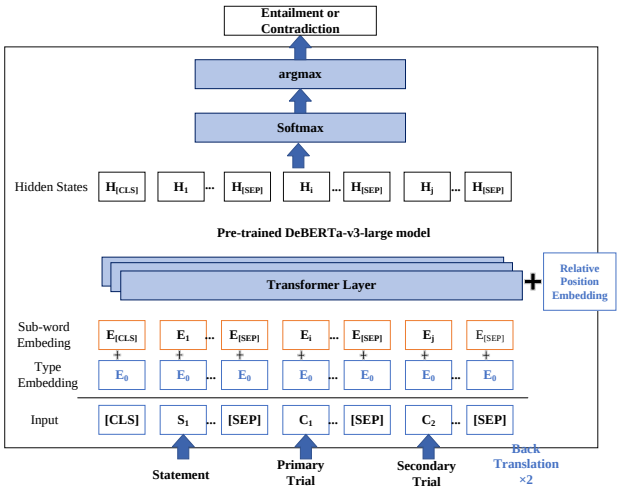

Figure 2: The structure of the system

with leading generative approaches, demonstrating that enhancements in model size correlate with performance improvements. Specifically, augmenting the model's scale directly boosts performance, significantly surpassing the gains from biomedical pre-training. Thus, validating the development set, we opted to submit results with higher experimental scores. Our approach involved fine-tuning the pre-trained DeBERTa-v3-large model, supplemented with supervised contrastive learning and back-translation techniques.

Comprehensive experiments showed that our system achieved a maximum F1 score of 0.77, securing the seventh position on the leaderboard. However, the model exhibited suboptimal performance in faithfulness and consistency metrics, indicating a weaker predictive capacity for data altered by interventions, highlighting areas for future enhancement.

The remainder of this paper is organized as follows. Section 2 describes the model and method used in our system, Section 3 discusses the results of the experiments, and finally, the conclusions are drawn in Section 4.

2 System Description

This section will describe the architecture of the proposed model in detail, including the data loader and back translation, the pre-trained model DeBERTa-v3-large, and supervised contrastive learning; the system model we proposed is shown in Figure 2.

2.1 Data preprocessing

Before feeding statements and CTRs into the model, preprocessing is performed. Initially, data augmentation is conducted through back-translation, a widely adopted technique involving translating text into another language and then back to the original language. This process, achieved via automatic translation systems, utilized Baidu's machine translation API[1] in this study, effectively doubling the training data. Given training data $\mathcal{D} = \{S, C, y\}$, y is the corresponding ground-true label, S is the medical hypothesis sentences, C is the corresponding CTR of the sentence, data loader is applied to transform training data as:

$$X = [CLS]s_1 s_2 \dots s_n[SEP]c_1 c_2 \dots c_m[SEP] \quad (1)$$

where s is the hypotheses with length n and c denotes the CTR reports with length m. [CLS] is a special mark indicating the beginning of the text sequence; [SEP] indicates the separator between text sequences. A similar process compares two CTRs, appending [SEP] and concluding similarly. Sequences exceeding 512 tokens are truncated, while shorter ones are padded.

2.2 Pre-trained DeBERTa-v3-large model

Given the commendable performance exhibited by the DeBERTa-v3 model (He et al., 2023) on this task (Jullien et al., 2023b) and the positive correlation between model parameter size and performance, the DeBERTa-v3-large model was selected as the baseline. Furthermore, an exploration was conducted with several DeBERTa-v3-large models fine-tuned on other NLI datasets available on the Hugging face[2] (Sileo, 2023; Laurer et al., 2023). The pre-trained datasets include MultiNLI, FeverNLI, ANLI, LingNLI, and WANLI. The DeBERTa-v3-large model has 24 layers and a hidden size of 1024. It has 304M backbone parameters with a vocabulary containing 128K tokens, which introduces 131M parameters in the Embedding layer.DeBERTa encodes the input text into the logits,

$$\mathbf{H} = \mathrm{Enc}(X; \theta) \quad (2)$$

where $\mathbf{H} \in \mathbb{R}^d$ is the logits with the dimensionality of d. The [CLS] token, positioned at the beginning of the input sequence, yields a hidden representation $\mathbf{H}_0$, signifying the sequence's initial contex-

[1] https://api.fanyi.baidu.com/
[2] https://huggingface.co/

tual semantic feature within the vector H. Following the acquisition of Following the acquisition of $\mathbf{H}_0$ the [CLS] representation, a fully connected layer leverages it to predict the corresponding label for the input text. The output is a softmax function,

$$\hat{\mathbf{y}} = \text{softmax}(W^0\mathbf{H}_0 + (h^0) \tag{3}$$

where $W^0 \in R^{d\times k}$ represents the weight of the fully connected layer, h^0 represents the offset of the fully connected layer, and k represents the number of classification labels.

2.3 Supervised Contrastive Learning Loss

Contrastive learning (Khosla et al., 2020)is a technique that learns to embed representations of similar samples closer together in the embedding space while pushing apart representations of dissimilar samples. In our model training, we employed this approach by incorporating a supervised contrastive loss alongside the cross-entropy loss. We hypothesized that this method would effectively handle interventions because it encourages the model to learn invariant features across different variations of the data introduced by such interventions (Feng et al., 2023). This invariance is critical for the model to generalize well to new, unseen data that might contain similar variations. Furthermore, we experimented with the R-drop technique R-drop(liang et al., 2021) to further enhance the model's generalization capabilities. However, results from Section 3 suggest that our implementation did not yield the expected improvements. This could be attributed to suboptimal parameter settings or the specific characteristics of our dataset and model size, which might have led to underfitting.

The cross-entropy loss is employed to guide the model towards accurate classification, which measures the discrepancy between the probability distribution predicted by the model and the actual distribution of the proper labels. The contrastive loss part h_i represents a feature vector, and h_{i+} is another feature vector within the same category. The dot product operation effectively calculates the cosine similarity between normalized feature vectors, τ which is the temperature parameter that modulates the model's ability to differentiate between pairs of samples. As the temperature parameter increases, the contrastive loss tends towards treating all sample pairs equally. In contrast, decreasing the temperature parameter focuses the model's attention on the most challenging negative samples. The indicator function ensures that a sample is not com-pared with itself. The SCL loss aims to bring samples of the same category closer together while pushing samples from different categories apart, thereby enhancing the discriminative power of the features. α and β hyperparameters are used to balance the contribution of each loss component. Ultimately, we formulated our loss function as follows to combine both losses effectively:

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{4}$$

$$L_{SCL} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(h_i \cdot h_{i+}/\tau)}{\sum_{j=1}^{N} 1_{[j+i]} \exp(h_i \cdot h_j/\tau)} \tag{5}$$

$$L = \alpha L_{CE} + \beta L_{SCL} \tag{6}$$

3 Experimental Results

Datasets. NLI4CT (Jullien et al., 2023a) is designed to assist in developing and benchmarking models for clinical NLI. Which consists of annotated Clinical Trial Reports (CTRs) focused on breast cancer research. Each CTR is meticulously structured into four key sections: (1)Eligibility Criteria: Specifies the prerequisites for patient inclusion in the clinical trial, detailing necessary conditions and characteristics. (2)Intervention: Describes the treatment regimen, including type, dosage, frequency, and duration of the administered treatments. (3)Results: Reports on the trial's participant count, outcome measures, metrics, and findings. (4)Adverse Events: Documents observed signs, symptoms, and any adverse effects encountered by patients during the clinical trial's course. The premises for NLI4CT are sourced from 1,000 publicly accessible Breast Cancer Clinical Trial Reports (CTRs) in English, published on ClinicalTrials.gov[3]. There are 999 breast cancer CTRs in the dataset. The datasets, which are divided into train, development, and test sets, contain a total of 2400 annotated statements. The distribution of labels be-tween the train and development sets is even. Upon employing back-translation, the volume of training data was effectively doubled. Notwithstanding, the test dataset substantially exceeds the size of the training dataset, a scenario that underscores the critical need for models to exhibit

[3]https://clinicaltrials.gov/

Class	Training	Validation	Enhancement Training	Test
Contradiction	850	100	1700	
Entailment	850	100	1700	
Total	1700	200	3400	5667

Table 1: Data distribution

robust generalization capabilities. The distribution of the dacctaset is shown in Table 1.

Evaluation Metrics. The task has three metrics; the **Macro F1-score** is a foundational metric, offering a balanced measure of precision and recall across the dataset's categorial spectrum without any semantic interventions. **Faithfulness** quantifies a model's capacity to adjust its predictions for the right reasons, especially when confronted with semantic-altering interventions. This metric illuminates a model's understanding of the underlying semantics, rewarding models that exhibit agile adaptability to semantic nuances. **Consistency** gauges a model's reliability in producing uniform outputs for semantically equivalent stimuli, regardless of the correctness of the final prediction. This metric champions models that demonstrate robustness in semantic representation, ensuring that semantically similar inputs yield consistent predictions. The formula for the three indices is expressed as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

$$Faithfulness = \frac{1}{N} \sum_{i=1}^{N} |f(y_i) - f(x_i)| \quad (8)$$

where $x_i \in C$ with $\text{Label}(x_i) \neq \text{Label}(y_i)$ and $f(y_i) = \text{Label}(y_i)$.

$$Consistency = \frac{1}{N} \sum_{i=1}^{N} 1 - |f(y_i) - f(x_i)| \quad (9)$$

where $x_i \in C$ where $\text{Label}(x_i) = \text{Label}(y_i)$, N is the total number of sentences, x_i and $y_i$denote the modified and original statements, respectively. The F1 score primarily aims to evaluate the model's performance on data without interventions. At the same time, the other two metrics assess the ability to make correct judgments post-intervention, indicating the model's deeper and more logical understanding of semantic information.

Implementation Details. All compared models were downloaded from HuggingFace. We fine-tune these models on the training set. The models

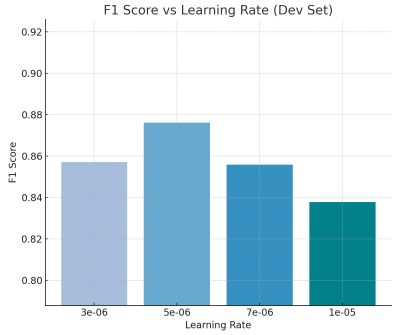

Figure 3: F1 scores on the development set for different learning rates, using the same pre-trained model and other parameters

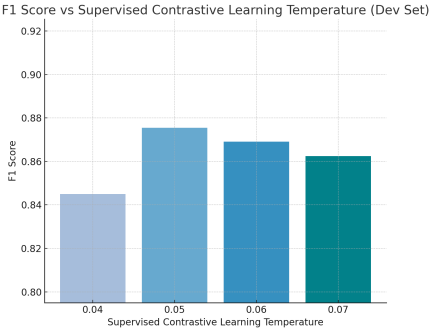

Figure 4: F1 scores on the development set for Supervised Contrastive Learning Temperature, using the same pre-trained model and other parameters

are evaluated on the validation every ten steps using precision, recall, and F1 scores. An Adam optimizer (Loshchilov and Hutter, 2019) updates the param-eters. The warmup strategy (He et al., 2016)is used to optimize the learning rate, and a fixed ran-dom seed is used.

Parameters Fine-tuning. Initially, manual adjustments were made to the hyperparameters, including the learning rate and the temperature for the contrastive loss function. Due to constraints imposed by GPU memory capacity, the batch size for training data was fixed at 4, with results illustrated in Figures 3 and 4. Upon identifying the approximate range of optimal parameters, the Optuna framework (Akiba et al., 2019) was employed for hyperparameter tuning. The parameters yielding the highest F1 score on the development set were selected for further training and model saving. The inference results were then uploaded to the platform.

Comparative Results and Discussion. Table 2 demonstrates that models pre-trained on additional datasets surpass the baseline model in performance on the development set. Nonetheless, it is shown

Model	Pre-training data	Loss	F1
Deberta-v3-large		CE	0.8018
Deberta-v3-large	600+ tasks	CE	0.8518
Deberta-v3-large	MultiNLI+FeverNLI+ANLI+LingNLI+WANLI	CE	0.8504
Deberta-v3-large	MultiNLI+FeverNLI+ANLI+LingNLI+WANLI+Other classification tasks	CE	0.8173
Deberta-v3-large	600+ tasks	CE+R-drop	0.8487
Deberta-v3-large	600+ tasks	CE+SCL	0.8544
Deberta-v3-large +Back Translation	600+ tasks	CE+SCL	0.8625
Deberta-v3-large +Back Translation	MultiNLI+FeverNLI+ANLI+LingNLI+WANLI	CE+SCL	0.8834
Deberta-v3-large +Back Translation	MultiNLI+FeverNLI+ANLI+LingNLI+WANLI+Other classification tasks	CE+SCL	0.8755

Table 2: Comparative results of experiments in the dev set

F1(dev)	F1(test)	Faithfulness	Consistency
0.8755	0.77	0.67	0.72
0.8834	0.75	0.73	0.74

Table 3: Optimal results of the test

that an excess of pre-training tasks yields minimal enhancements in model performance, such as the model that was fine-tuned with multi-task learning across over 600 tasks from the task source collection (Sileo, 2023; Laurer et al., 2023). It was also observed that R-drop might not be ideally suited for this task, potentially due to suboptimal parameter selection. It can be seen from Figure 3 and Figure 4 that the learning rate of the model is more suitable in the vicinity of 5e-6.

In contrast, the temperature of comparative learning is difficult to control, and the model performance is not linear, which needs further exploration. A degree of performance improvement was achieved through supervised contrastive learning. The highest F1 score of 0.8834 on the development set was achieved by combining supervised contrastive learning with the back-translation method. However, an F1 score of 0.75 was only achieved by this model on the test set, equating to the score of 11th place. Scores of 0.73 and 0.74 were reached on the other two metrics, comparable to the scores of the 17th and 9th places, respectively. Despite this, only the highest F1 scores are listed on the leaderboard. Another model of ours reached an F1 score of 0.77, placing it 9th, yet the scores on the other two metrics were not as high, placing 17th and 13th, respectively.

Such scores suggest that predictions are often not based on valid reasoning by the model. Accurate conclusions, when reached, may be derived from incorrect premises or misinterpretations of the input data, suggesting an insensitivity to semantic changes or an incapacity to reflect these changes accurately in its predictions. The reduction of this score indicates the model's prediction instability in the absence of significant semantic alterations, reflecting an excessive sensitivity to minor variations in input or a failure to capture and maintain the input's core semantic features accurately.

These findings reveal deficiencies in our system's ability to understand and process complex and subtle semantic changes despite adequate performance, as indicated by the F1 score. An overreliance on specific data distributions, a lack of generalizability, or challenges in explaining decisions in practical applications may result. To improve the model's Faithfulness and Consistency, it may be necessary for further research and improvements to be conducted on the model's internal representations and training processes or for additional mechanisms to be integrated for better processing of semantic information.

4 Conclusion

This paper introduces a system based on fine-tuning and pre-training Deberta-v3-large for SemEval2024 task 2, targeting safe biomedical NLI for clinical trials. Achieving seventh out of 32 with an F1 of 0.77 showcases the effectiveness of multi-task pre-training, supervised contrastive learning, and back-translation despite struggles with intervention data and deep semantic under-standing. Issues include truncated evidence from extended clinical trial premises (Kong et al., 2022)and insufficient model depth for causal reasoning. Future research should enhance semantic comprehension and causal reasoning and refine contrastive learning to improve the handling complex data and interventions, aiming to overcome current limitations in safe biomedical NLI.

5 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant

Nos. 61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902.

Nancy E Avis, Kevin W Smith, Carol L Link, Gabriel N Hortobagyi, and Edgardo Rivera. 2006. Factors associated with participation in breast cancer treatment clinical trials. *J Clin Oncol*, 24(12):1860–1867.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Chao Feng, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at SemEval-2023 task7: Multi-evidence natural language inference for clinical trial data based a BioBERT model. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 664–670, Toronto, Canada. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

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. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical bert with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building efficient universal classifiers with natural language inference. *ArXiv*, abs/2312.17543.

xiaobo liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905. Curran Associates, Inc.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Damien Sileo. 2023. tasksource: A dataset harmonization framework for streamlined nlp multi-task learning and evaluation. *ArXiv*, abs/2301.05948.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

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. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1863–1870, Toronto, Canada. Association for Computational Linguistics.

Mengyuan Zhang, Jin Wang, and Xuejie Zhang. 2020. Using a pre-trained language model for medical named entity extraction in chinese clinic text. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 312–317.

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. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1681–1690, Toronto, Canada. Association for Computational Linguistics.