NCUEE-NLP at SemEval-2023 Task 7: Ensemble Biomedical LinkBERT Transformers in Multi-evidence Natural Language Inference for Clinical Trial Data

Chao-Yi Chen, Kao-Yuan Tien, Yuan-Hao Cheng, and Lung-Hao Lee
Department of Electrical Engineering
National Central University
No. 300, Zongda Rd., Zhongli Dist., Taoyuan City 32001, Taiwan
{110581007, 110521083, 110521079}@cc.ncu.edu.tw, lhlee@ee.ncu.edu.tw

Abstract

This study describes the model design of the NCUEE-NLP system for the SemEval-2023 NLI4CT task that focuses on multi-evidence natural language inference for clinical trial data. We use the LinkBERT transformer in the biomedical domain (denoted as BioLinkBERT) as our main system architecture. First, a set of sentences in clinical trial reports is extracted as evidence for premise-statement inference. This identified evidence is then used to determine the inference relation (i.e., entailment or contradiction). Finally, a soft voting ensemble mechanism is applied to enhance the system performance. For Subtask 1 on textual entailment, our best submission had an F1-score of 0.7091, ranking sixth among all 30 participating teams. For Subtask 2 on evidence retrieval, our best result obtained an F1-score of 0.7940, ranking ninth of 19 submissions.

1 Introduction

Natural Language Inference (NLI) seeks to determine whether a given hypothesis is true (i.e., entailment), false (contradiction), or undetermined (neutral) according to a known premise. Methods used for NLI task range from earlier symbolic and statistical approaches to recent deep-learning-based models (Storks et al., 2019). The BERT-BiLSTM-Attention model was proposed for medical text inference (Lee et al., 2019). A knowledge adaptive approach was derived to encode the premise/hypothesis text for improving medical NLI (Chowdhury et al., 2020). Structured domain knowledge from the Unified Medical Language System (UMLS) was incorporated into the ESIM model (Chen et al., 2017) to improve NLI performance in the medical domain (Sharma et al., 2019). Clinical domain knowledge of BERT models (Devlin et al., 2019) was explored using lexical retrieval, syntactic retrieval, and classification models for external knowledge integration (Sushil et al., 2021).

The SemEval-2023 Task 7 focuses on multi-evidence natural language inference for clinical trial data (denoted as NLI4CT) (Jullien et al., 2023). The NLI4CT task is mainly based on a collection of Clinical Trial Reports (CTR) for breast cancer, with statements, explanations, and labels annotated by domain expert annotators. It includes two subtasks: 1) *Textual Entailment*: determining the inference relation (i.e., entailment or contradiction) between CTR-statement pairs; 2) *Evidence Retrieval*: extracting a set of supporting facts in a given CTR premise to justify the label predicted in Subtask 1.

The MEDIQA-2019 shared task covers an NLI subtask in the medical domain (Ben Abacha et al., 2019), in which most systems built on the BERT model are pre-trained on a large-scale open domain corpus and its variants focus on domain-specific knowledge such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), and ClinicalBERT (Alsentzer et al., 2019). Given the promising results obtained by most such approaches, we apply BERT-like neural networks to the NLI task in the clinical domain.

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical **E**ngineering, **N**atural **L**anguage **P**rocessing Lab) system for SemEval-2023 NLI4CT task. We use biomedical LinkBERT (Yasunaga et al., 2022) as our main system architecture to first extract a set of

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 776–781

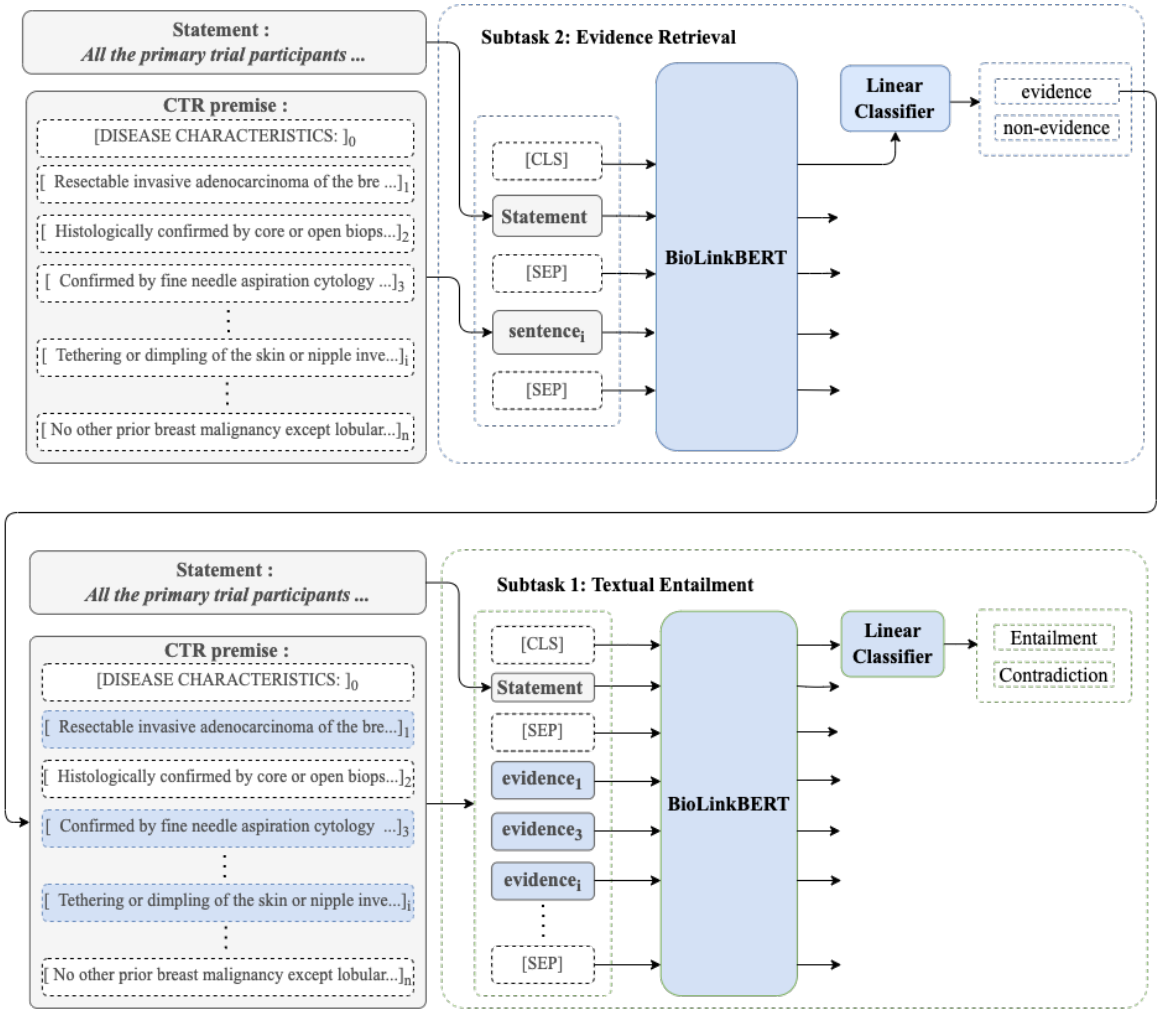

Figure 1: Our NCUEE-NLP system architecture for the NLI4CT task.

supporting facts (i.e., for Subtask 2), and then determine the inference relation (for Subtask 1) in a given CTR-statement pair. A soft voting ensemble mechanism was used to sum the predicted probability for class labels from biomedical LinkBERT transformers fine-tuned on different publicly available data sets for the NLI task. Finally, the class label with the largest summed probability will be precited for system performance evaluation. For Subtask 1, our best submission with an F1-score of 0.7091 ranked sixth among all 30 participating teams. For Subtask 2, our best result had an F1-score of 0.7940, ranking ninth of 19 submissions.

The rest of this paper is organized as follows. Section 2 describes the NCUEE-NLP system for the NLI4CT task. Section 3 presents the results and performance comparisons. Conclusions are finally drawn in Section 4.

2 The NCUEE-NLP System

Figure 1 shows our NCUEE-NLP system architecture for the NLI4CT task. Our system is composed of two main parts: evidence retrieval for Subtask 2 and textual entailment for Subtask 1. Both subtasks mainly depend on the LinkBERT transformer (Yasunaga et al., 2022). LinkBERT is an improved version of BERT that captures documents links such as hyperlinks and citation links to include knowledge that spans across multiple documents. In addition to BERT pretrained on single documents, LinkBERT is pretrained by feeding linked documents into the same language model context. We use the biomedical LinkBERT (denoted as BioLinBERT) pretrained on PubMed with citation links for both subtasks.

For evidence retrieval (Subtask 2), a CTR premise is represented as a sequence of described sentences. Each sentence and its corresponding statement were grouped to train the BioLinkBERT as a classification problem. If a sentence-statement pair is true (i.e., evidence), the class is assigned as 1, and 0 otherwise (non-evidence). To classify a sentence-statement pair during the test phase, we use the output probability of the [CLS] token as an indicator for classification. The class with the highest probability will be regarded as the inference result.

For Subtask 1 on textual entailment, the identified evidence in the previous phase will be concatenated along with corresponding statements for label classification. If evidence sentences support the statement annotated with the entailment label, the class is assigned as 1, and 0 otherwise (contradiction). Similarly, the class with highest probability of the [CLS] token will be outputted as prediction result.

In addition, a soft voting mechanism, which involves summing the predicted probabilities for class labels and predicting the class label with the largest sum probability, was used to integrate different fine-tuned versions of the pretrained BioLinkBERT transformer for better classification performance.

3 Experiments and Results

3.1 Data

The datasets were mainly provided by task organizers (Jullien et al., 2023). The collected breast cancer CTRs were summarized into the following four sections: 1) *Eligibility criteria*: a set of patient conditions to be allowed for inclusion in the clinical trial; 2) *Intervention*: a set of information regarding to the type, dosage, frequency, and duration of treatments; 3) *Results*: numbers of participants in the trial, outcome measures, units and the results; and 4) *Adverse events*: observed signs and symptoms in patients during the clinical trial. For Subtask 1 on textual entailment, the annotated statements denoted in the set of sentences were used to make the inference decision (i.e., entailment or contradiction) of claims against sections in a single CTR premise or two compared CTRs. For Subtask 2 on evidence retrieval, given a CTR premise and a statement, the system should indicate which sentences in the premise can be regarded as evidence to support the label predicted in Subtask 1.

A total of 1700 CTRs with annotations in the training set were used to fine-tune the pretrained transformer models. During the development phase, 200 annotated CTRs in the development set were used to develop the system and obtain optimized parameters. Finally, the test set containing 500 CTRs with the corresponding annotations was used to evaluate the system performance for both subtasks.

3.2 Settings

We used the BioLinkBERT-large model (Yasunaga et al., 2022), which was pretrained on PubMed abstracts along with citation link information and then fine-tuned on different datasets for this NLI4CT task. Three NLI datasets were used to fine-tune the model, including 1) MedNLI (Romanov and Shivade, 2018): this contains patient medical histories annotated by doctors to perform NLI tasks in the clinical domain; 2) MultiNLI (Wang et al., 2018): this is a crowdsourced collection of multi-genre sentence pairs with textual entailment annotations in the GLUE benchmark data; and 3) SNLI (Bowman et al., 2015): the Stanford NLI corpus is a collection of human-written sentence pairs manually labeled to recognize textual entailment.

For performance comparison, we used the BioBERT (Lee et al., 2020), which was pretrained on PubMed abstracts and PubMed Central full texts (denoted as PMC). The discharge summaries from the MIMIC III database (Johnson et al., 2016) or the whole database were then used to fine-tune the model (Alsentzer et al., 2019).

All compared models were downloaded from HuggingFace[1]. We continuously fine-tuned these models using the training dataset. The hyperparameter values were optimized as follows: embedding size 512; epochs 100 with early stopping mechanism; batch size 8; learning rate 7e-6 for Subtask 1 and 5e-6 for Subtask 2.

[1] https://huggingface.co/cnut1648/biolinkbert-mednli
https://huggingface.co/cnut1648/biolinkbert-mnli
https://huggingface.co/cnut1648/biolinkbert-large-mnli-snli
https://huggingface.co/dmis-lab/biobert-v1.1
https://github.com/EmilyAlsentzer/clinicalBERT

Model		Subtask 1 Textual Entailment			Subtask 2 Evidence Retrieval		
Transformer (Pre-trained data)	Fine-tuned data	Pre.	Rec.	F1	Pre.	Rec.	F1
BioBERT (PubMed + PMC)	-	0.6031	0.7900	0.6753	0.7777	0.8589	0.8163
	MIMIC III	0.5942	0.8200	0.6891	0.7853	0.8730	0.8286
	Discharge summaries	0.5804	0.8300	0.7102	0.8069	0.8179	0.8124
Ensemble		0.6015	0.8000	0.6867	0.7950	0.8604	0.8264
BioLinkBERT (PubMed)	MedNLI	0.7094	0.8300	**0.7650**	0.8316	0.8599	0.8455
	MultiNLI	0.6393	0.7800	0.7111	0.7778	0.8982	0.8337
	MultiNLI + SNLI	0.6719	0.8600	0.7574	0.7921	0.8877	0.8372
Ensemble		0.6563	0.8400	0.7368	0.8093	0.8951	**0.8500**

Table 1: Biomedical transformer results on the development data.

Model		Subtask 1 Textual Entailment			Subtask 2 Evidence Retrieval		
Transformer (Pre-trained data)	Fine-tuned data	Pre.	Rec.	F1	Pre.	Rec.	F1
BioBERT (PubMed + PMC)	-	0.6032	0.7480	0.6678	0.7823	0.7471	0.7643
	MIMIC III	0.5899	0.6560	0.6212	0.7705	0.7936	0.7819
	Discharge summaries	0.5718	0.7800	0.6598	0.7882	0.7378	0.7622
Ensemble		0.6197	0.7040	0.6592	0.7921	0.7719	0.7819
BioLinkBERT (PubMed)	MedNLI	0.6907	0.6880	0.6893	0.7914	0.7542	0.7724
	MultiNLI	0.6523	0.7280	0.6880	0.7827	0.8037	0.7931
	MultiNLI + SNLI	0.6258	0.7760	0.6928	0.8027	0.7675	0.7847
Ensemble		0.6678	0.7560	**0.7091**	0.8028	0.7855	**0.7940**

Table 2: Biomedical transformer results on the test data.

The evaluation metrics of this shared task are standard precision, recall, and F1-score for both subtasks. The maximum submission limitations for Subtasks 1 and 2 were respectively 100 and 30. The final ranking is determined from the best submission based on macro-averaging F1-score.

3.3 Results

Table 1 shows the results of our submissions on the development set. BioLinkBERT clearly outperformed BioBERT for both subtasks for all fine-tuned BioLinkBERT versions. For Subtask 1 on textual entailment, BioLinkBERT fine-tuning on MedNLI achieved outperformed those versions on other NLI datasets and the ensemble method. This indicates that the medical NLI dataset is more suitable for this shared task in the clinical domain. For Subtask 2 on evidence retrieval, the fine-tuned MedNLI version had the second-best result. The ensemble mechanisms on these three fine-tuned versions improved the performance resulting in the best result.

Tables 2 shows the results of our submissions on the test set. Ensemble BioLinkBERT transformer outperformed other models for both subtasks, confirming the effectiveness of our ensemble mechanism. For the textual entailment subtask, our best submission achieved an F1-score of 0.7091,

and ranked sixth among all 30 participating teams. For the evidence retrieval subtask, our best result obtained an F1-score of 0.7940, ranking ninth of 19 submissions.

4 Conclusions

This study describes the NCUEE-NLP system in the SemEval-2023 NLI4CT task, including system design, implementation and evaluation. We integrate different fine-tuned versions of BioLinkBERT model to retrieve evidence in the CTR premise for textual entailment in a given statement. For Subtask 1, our best submission obtained an F1-score of 0.7091, ranking sixth among all 30 participating teams. For Subtask 2, our best result obtained an F1-score of 0.7940, ranking ninth of 19 submissions.

Acknowledgments

This study is partially supported by the National Science and Technology Council, Taiwan, under the grant MOST 111-2628-E-008-005-MY3.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th Biomedical Natural Language Processing Workshop and Shared Task*, Association for Computational Linguistics, pages 370–379. http://dx.doi.org/10.18653/v1/W19-5039

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, pages 72-78. http://dx.doi.org/10.18653/v1/W19-1909

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pages 3615-3620. http://dx.doi.org/10.18653/v1/D19-1371

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*. Association for Computational Linguistics, pages 632-642. http://dx.doi.org/10.18653/v1/D15-1075

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1657-1668. http://dx.doi.org/10.18653/v1/P17-1152

Shaika Chowdhury, Philip Yu and Yuan Luo. 2020. Improving medical NLI using context-aware domain knowledge. In *Proceedings of the 9th Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, pages 1-11.

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)*. Association for Computational Linguistics, pages 4171-4186. http://dx.doi.org/10.18653/v1/N19-1423

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, Article 160035.

Mael Jullien, Marco Valentino, Hannah Frost, and Paul O'Regan, Donal Landers, and André Freitas. 2023. SemEval-2023 Task 7: multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Lung-Hao Lee, Yi Lu, Po-Han Chen, Po-Lei Lee, and Kou-Kai Shyu. NCUEE at MEDIQA 2019: medical text inference using ensemble BERT-BiLSTM-Attention model. In *Proceedings of the 18th Biomedical Natural Language Processing Workshop and Shared Task*. Association for Computational Linguistics, pages 528-532. http://dx.doi.org/10.18653/v1/W19-5058

Alexey Romanov, and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1586-1596. https://www.aclweb.org/anthology/D18-1187

Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pages 6092-6097. http://dx.doi.org/10.18653/v1/D19-1631

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Recent advances in natural language inferences: a survey of benchmarks, resources, and approaches. *arXiv Preprint*, arXiv:1904.01172v3. https://doi.org/10.48550/arXiv.1904.01172

Madhumita Sushil, Simon Suster, and Walter Daelemans. 2021. Are we there yet? Exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, pages 41-53. http://dx.doi.org/10.18653/v1/2021.bionlp-1.5

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, pages 353-355. http://dx.doi.org/10.18653/v1/W18-5446

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 8003-8016. http://dx.doi.org/10.18653/v1/2022.acl-long.551