NCUEE-NLP at SemEval-2023 Task 8: Identifying Medical Causal Claims and Extracting PIO Frames Using the Transformer Models

Lung-Hao Lee, Yuan-Hao Cheng, Jen-Hao Yang, and Kao-Yuan Tien
Department of Electrical Engineering
National Central University
No. 300, Zongda Rd., Zhongli Dist., Taoyuan City 32001, Taiwan
lhlee@ee.ncu.edu.tw, {110521079, 111521054, 110521083}@cc.ncu.edu.tw

Abstract

This study describes the model design of the NCUEE-NLP system for the SemEval-2023 Task 8. We use the pre-trained transformer models and fine-tune the task datasets to identify medical causal claims and extract population, intervention, and outcome elements in a Reddit post when a claim is given. Our best system submission for the causal claim identification subtask achieved a F1-score of 70.15%. Our best submission for the PIO frame extraction subtask achieved F1-scores of 37.78% for Population class, 43.58% for Intervention class, and 30.67% for Outcome class, resulting in a macro-averaging F1-score of 37.34%. Our system evaluation results ranked second position among all participating teams.

1 Introduction

A medical causal claim is an assertion that invokes causal relationships between variables such as "a drug has certain effects on a disease" (Pearl, 2004). A manually-annotated corpus for causal statements in PubMed publications has been used to train prediction models for identifying "correlational," "conditional causal," "direct causal," and "no relationship" statements in research conclusion sections (Yu et al., 2019). BioBERT (Lee et al., 2020) has been shown to outperform BERT (Devlin et al., 2019) and linear SVM in identifying exaggerated causal claims made in health press releases (Yu et al., 2020). Empirical analyses have found that biomedical tweets are densely populated with claims (Wührl and Klinger, 2021). Entity-based claim representations have been proposed to extract condensed claims from medical tweets (Wührl and Klinger, 2022).

The PICO framework is widely used to identify sentences in a given medical text belonging to four elements: 1) Population (P): this usually describes the characteristics of populations involved; 2) Intervention (I): this addresses the primary intervention along with any risk factors; 3) Comparison (C): this compares the efficacy of any new interventions with the primary intervention; and 4) Outcome (O): this measures the results of the intervention including improvements or side effects. The PICO elements have been used as a knowledge representation framework to analyze clinical questions (Huang et al., 2006). A corpus of clinical trial publications has been annotated with PICO elements (Zlabinger et al., 2018). PICO elements in medical texts have been detected based on long short-term memory networks (Jin and Szolovits, 2018). The contextualized BERT embedding has been used to enhance PIO element detection performance (Mezaoui et al., 2019).

The SemEval-2023 Task 8 (Khetan, et al., 2023) organized a challenge to identify medical causal claims and extract related PIO frames. This task consists of two subtasks. 1) Subtask 1: for the provided snippet of text, participants should identify the span of text that belongs to either four defined categories, including "*Claim*," "*Experience*," "*Experience based claim*," and "*Question*". 2) Subtask 2: for the given text snippet and its identified claim, participants should extract the related "*Population*", "*Intervention*," and "*Outcome*" frames.

This paper describes the **NCUEE-NLP** (**N**ational **C**entral **U**niversity, Dept. of **E**lectrical

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 312–317
July 13-14, 2023 ©2023 Association for Computational Linguistics

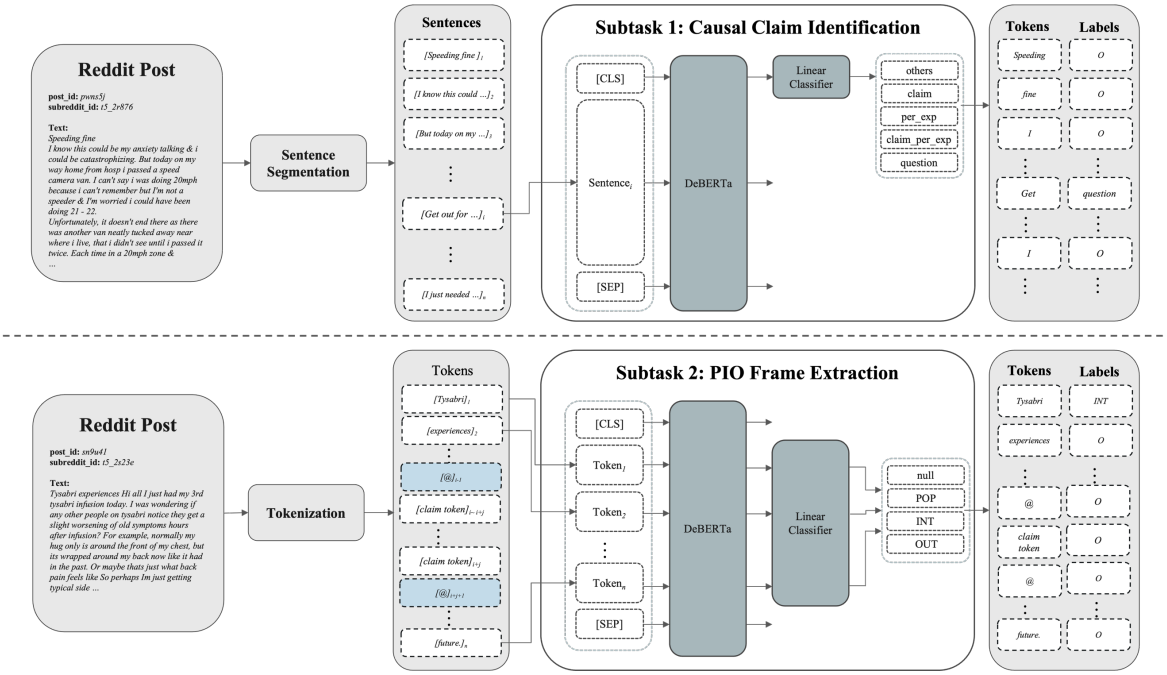

Figure 1: Our NCUEE-NLP system workflow for SemEval-2023 Task 8.

Engineering, **N**atural **L**anguage **P**rocessing Lab) system for the SemEval-2023 Task 8 (Khetan, et al., 2023). We explore transformer-based neural computing models to identify medical causal claims and extract their PIO frames. Based on experimental results using datasets provided by task organizers, we find the DeBERTa (He et al., 2021a; 2021b) transformers outperformed other models in 5-fold cross validation of the training set, but achieved second-best results for performance evaluation on the test set. Finally, our best submissions ranked second in the leaderboard for both subtasks.

The rest of this paper is organized as follows. Section 2 describes our developed NCUEE-NLP system for the SemEval-2023 Task 8. Section 3 presents the evaluation results and performance comparisons. Conclusions are finally drawn in Section 4.

2 The NCUEE-NLP System

Figure 1 shows the system workflow. Our model mainly depends on the DeBERTa transformer (He et al., 2021a). Decoding-enhanced BERT with disentangled attention (DeBERTa) improves the BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models using two novel techniques: 1) a disentangled attention mechanism, and 2) an enhanced mask decoder. For disentangled attention computation, each word is represented using content and position vectors, and then attention weights among words are computed by applying disentangled matrices to their contents and relative positions. In the enhanced mask decoder architecture, absolute positions in the decoding layer are used to predict the masked tokens in model pre-training. DeBERTaV3 (He et al., 2021b) further improves the original DeBERTa model by replacing the masked language learning (MLM) pre-training task with an ELECTRA-style replaced token detection (RTD) task (Clark et al., 2020). In addition, a new gradient-disentangled embedding sharing method is proposed to avoid the tug-of-war dynamics for improving both the training efficiency and the quality of the original DeBERTa pre-trained model.

For Subtask 1, we fine-tune the pre-trained DeBERTa transformer as a sentence-level classification task. First, we segment each post's content into several sentences using Trankit (Nguyen et al., 2021), a light-weight transformer-based toolkit for multilingual NLP tasks. Each sentence is then assigned to one of the following defined categories: 1) *Claim*: communicating a causal interaction between an intervention and an

outcome; 2) *Experience*: relating a specific outcome to an intervention or population based on someone's experience; 3) *Experience based claim*: a claim based on someone's experience; 4) *Question*: posing a question; and 5) *Others*: sentences which do not belong to the above categories. A sentence may be composed of several tokens with different annotated categories based on the datasets provided by task organizers. We determine a sentence category using the majority voting mechanism from those constituent tokens. During the evaluation phase, we then convert an obtained sentence prediction result into token-level labels for performance evaluation. No separate fine-tuning of classification at the token level is needed. Each sentence label will be assigned to all of its tokens.

For Subtask 2, we fine-tune the pre-trained DeBERTa transformer with a weak supervised mechanism (Huang et al., 2019) as a sequence labeling task. The weak supervised signal "@" is used to emphasize the starting and ending positions in a given claim. A sentence represented in terms of a token sequence will be aligned with Population (noted as Pop-tag), Intervention (Int-tag) and Outcome (Out-tag) for frame extraction. A null-tag is used to indicate that a token belongs to no extracted PIO frames. For performance evaluation, we transform the sequence labeling output to obtain the starting and ending positions of extracted spans corresponding to each PIO frame.

3 Experiments and Results

3.1 Data

The experimental datasets were mainly provided by the task organizers (Wadhwa et al., 2023). All datasets were built from Reddit posts. Due to privacy concerns, task organizers only provide Reddit post identifiers, annotations, and a script to obtain the data and merge it with the provided annotations. If the post is subsequently deleted by the user, the script won't be able to obtain the post content.

For Subtask 1, we used the script to obtain 5,287 posts, accounting for 88.6% among the 5,965 provided post identifiers in the training set. The remaining 678 posts (11.4%) had been deleted by users, including those cases in which the post text was only partially removed. Similarly, we obtained 1,264 posts (88.8%) from the 1,424 post identifiers in the test set, while 160 posts (11.2%) lack the whole or partial post content. Our obtained Reddit posts were segmented using the Trankit toolkit (Nguyen et al., 2021) to result in training and test sets respectively with 43,237 and 11,492 sentences.

For Subtask 2, both training and test datasets are small. We used the same script to obtain 524 posts (87.8% of a total 597) as the training data. The test set contained only 130 posts (86.7% of a total 150) for system performance evaluation.

3.2 Settings

In addition to DeBERTa (He et al., 2021a; 2021b), we also used BERT (Devlin et al., 2019), BioBERT (Lee et al., 2019), and RoBERTa (Liu et al., 2019) to compare the performance of different transformers. We downloaded these pre-trained models from the Huggingface[1] and fine-tuned the downstream tasks using the official training datasets for each task. We used the 5-fold cross-validation to compare performance during the system development phase. The hyper-parameter values were manually optimized as follows: batch size 2, epochs 20 with early stopping mechanism, AdamW optimizer and learning rate 1e-5 for Subtask 1 and 4e-5 for Subtask 2.

The evaluation metrics of this shared task are standard precision, recall and F1-score. The final ranking is determined from the best submission based on macro-averaging F1-score.

3.3 Results

Tables 1 shows the results of 5-fold cross-validation on the training set. For the causal claim identification subtask, BioBERT slightly improved on the performance achieved by BERT, which may be due to domain-specific data being pre-trained on the same BERT architecture. RoBERTa is an enhanced version of BERT that outperforms BioBERT and BERT. The DeBERTa transformer, achieved the best F1-score of 56.48%, outperforming RoBERTa and BERT through the use of two novel techniques. For the PIO frame extraction subtask, we obtained closely consistent

[1] https://huggingface.co/bert-large-uncased
https://huggingface.co/dmis-lab/biobert-large-cased-v1.1
https://huggingface.co/roberta-large
https://huggingface.co/microsoft/deberta-large
https://huggingface.co/microsoft/deberta-v3-large

Model		Subtask 1 Causal Claim Identification			Subtask 2 PIO Frame Extraction			
Transformer	version	Precision	Recall	F1	Pop-F1	Int-F1	Out-F1	Marco-F1
BERT	large uncased	53.67	54.26	53.81	36.71	32.59	24.93	31.41
BioBERT	large v1.1	56.13	53.49	53.93	44.07	39.18	25.14	36.13
RoBERTa	large	58.29	55.65	55.91	43.38	40.12	28.07	37.19
DeBERTa	large	**58.42**	**56.27**	**56.48**	44.45	38.03	26.03	36.17
	large v.3	57.01	55.61	55.48	**44.84**	**42.29**	**30.62**	**39.25**

Table 1: Transformers results on the training set (5-fold cross validation).

Model		Subtask 1 Causal Claim Identification			Subtask 2 PIO Frame Extraction			
Transformer	version	Precision	Recall	F1	Pop-F1	Int-F1	Out-F1	Marco-F1
BERT	large uncased	71.24	67.30	**70.15**	27.05	41.64	30.05	32.91
BioBERT	large v1.1	72.16	67.13	69.55	33.09	43.20	26.52	34.27
RoBERTa	large	70.90	67.17	68.98	**37.78**	43.58	**30.67**	**37.34**
DeBERTa	large	72.97	**67.36**	70.05	32.80	47.24	29.03	36.36
	large v.3	**73.33**	67.11	70.08	33.87	**50.72**	26.31	36.97

Table 2: Transformers results on the test set.

results. DeBERTa version 3 obtained the best macro-averaged F1-score of 39.25%.

Table 2 shows the test set results. We selected the best model with the highest F1-score for the 5-fold cross validation settings of each model to predict the testing instances for final evaluation submissions. DeBERTa version 3 achieved the second-best results of F1-scores for both subtasks. For the causal claim identification subtask, the original BERT performed slightly better (0.07%) than DeBERTa version 3. For the PIO frame extraction subtask, RoBERTa model achieved the best macro-averaging F1-score of 37.34%, although DeBERTa version 3 clearly had the best F1-score of 50.72% for the Intervention class.

In summary, for Subtask 1 on causal claim identification, our best submission had an F1-score of 70.15%, ranking second among a total of 7 participating teams. For Subtask 2 on PIO frame extraction, the class-wise F1-scores of our best submission were 37.78% for Population class, 43.58% for Intervention class, and 30.67% for Outcome class, resulting in a macro-averaged F1-score of 37.34%, each ranking second among a total of 6 participating teams.

4 Conclusions

This study describes the NCUEE-NLP system in the SemEval-2023 Task 8, including system design, implementation and evaluation. We used different versions of pre-trained transformer models and fine-tuned two subtasks using Reddit posts with annotations provided by the organizers. Our best submissions ranked second among all participating teams for all subtasks and evaluation metrics.

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

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International*

Conference on Learning Representations, pages 1-18. https://doi.org/10.48550/arXiv.2003.10555

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTa: decoding-enhanced BERT with disentangled attention. *arXiv Preprint*, arXiv:2006.03654v6. https://doi.org/10.48550/arXiv.2006.03654

Pengcheng He, Jianfeng Gao, Weizhu Chen. 2021b. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv Preprint*, arXiv:2111.09543v2. https://doi.org/10.48550/arXiv.2111.09543

Luyao Huang, Chi Sun, Xipeng Qiu, Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509-3514. http://dx.doi.org/10.18653/v1/D19-1355

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *Proceedings of the AMIA 2006 Annual Symposium*, pages 359-363.

Di Jin, and Peter Szolovits. 2018. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the 17th Workshop on Biomedical Language Processing*, Association for Computational Linguistics, pages 67-75. http://dx.doi.org/10.18653/v1/W18-2308

Vivek Khetan, Somin Wadhwa, Byron Wallace, and Silvio Amir. 2023. SemEval-2023 task 8: causal medical claim identification and related PIO frame extraction from social media posts. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv Preprint*, arXiv:1907.11692v1

Hichem Mezaoui, Aleksandr Gontcharov, and Isuru Gunasekara. 2019. Enhancing PIO element detection in medical text using contextualized embedding. In *Proceedings of the 18th Workshop on Biomedical Language Processing*, Association for Computational Linguistics, pages 217-222. http://dx.doi.org/10.18653/v1/W19-5023

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: a light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, pages 80-90. http://dx.doi.org/10.18653/v1/2021.eacl-demos.10

Judea Pearl. 2004. Robustness of causal claims. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligences*, Association for Computing Machinery, pp. 446-453. https://doi.org/10.48550/arXiv.1207.4173

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. RedHOT: a corpus of annotated medical questions, experiences, and claims on social media. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. https://arxiv.org/abs/2210.06331

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, Association for Computational Linguistics, pages 131-142. http://dx.doi.org/10.18653/v1/2021.bionlp-1.15

Amelie Wührl and Roman Klinger. 2022. Entity-based claim representation improves fact-checking of medical content in tweets. In *Proceedings of the 9th Workshop on Argument Mining*, Association for Computational Linguistics, pages 187-198.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, pages 4860-4872. http://dx.doi.org/10.18653/v1/2020.coling-main.427

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. 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 4664-4674. http://dx.doi.org/10.18653/v1/D19-1473

Markus Zlabinger, Linda Andersson, Allan Hanbury, Michael Andersson, Vanessa Quasnik, and Jon Brassey. 2018. Medical entity corpus with PICO elements and sentiment analysis. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, European Language Resources Association, pages 292-296.