USMBA-NLP at SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials using Bert

Anass Fahfouh[1], **Abdessamad Benlahbib**[1], **Jamal Riffi**[1] , **Hamid Tairi**[1]
[1] LISAC Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
anassfahfouh@gmail.com, abdessamad.benlahbib@usmba.ac.ma ,
riffi.jamal@gmail.com, htairi@yahoo.fr

Abstract

This paper presents the application of BERT in SemEval 2024 Task 2, Safe Biomedical Natural Language Inference for Clinical Trials. The main objectives of this task were: First, to investigate the consistency of BERT in its representation of semantic phenomena necessary for complex inference in clinical NLI settings. Second, to investigate the ability of BERT to perform faithful reasoning, i.e., make correct predictions for the correct reasons. The submitted model is fine-tuned on the NLI4CT dataset, which is enhanced with a novel contrast set, using binary cross entropy loss.

1 Introduction

NLI stands for Natural Language Inference. It is a task in natural language processing (NLP) where the goal is to determine the relationship between two text segments: a premise and a hypothesis. The task typically involves classifying whether the hypothesis is entailed, contradicted, or neutral with respect to the premise.

NLI has emerged as a beacon of hope for Clinical Trial Reports (CTRs). Its ability to handle vast amounts of medical evidence could revolutionize the interpretation and retrieval of CTRs. Clinical trials stand as pillars in experimental medicine, scrutinizing the efficacy and safety of novel treatments (Avis et al., 2006). CTRs meticulously outline trial methodologies and findings, guiding the development of targeted interventions for patients. Yet, the staggering quantity of published CTRs renders manual review impractical for devising new treatment protocols (DeYoung et al., 2020).

The proposed textual entailment task aims to advance the understanding of models' behavior and enhance existing evaluation methodologies for clinical NLI. By systematically applying controlled interventions, each engineered to probe a specific semantic phenomenon in natural language and numerical inference, the study seeks to assess the robustness, consistency, and faithfulness of models in clinical settings, thereby investigating their reasoning capabilities.

In this paper, we present our findings on SemEval 2024 Task 2, Safe Biomedical Natural Language Inference for Clinical Trials (Jullien et al., 2024). The aim of our method is to assess the robustness, consistency, and faithfulness of BERT (Devlin et al., 2019), Pre-training of Deep Bidirectional Transformers for Language Understanding, on the clinical NLI. Our method follows to steps: the first step is fine-tuning BERT on the Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) (Jullien et al., 2023) which is enhanced with a novel contrast set. Then, the prediction step, consists on the determining the inference relation (entailment vs contradiction) between CTR - statement pairs.

The rest of the paper is structured in the following manner: Section 2 provides the main objective of the Task. Section 3 describes our system. Section 4 details the experiments. And finally, Section 5 concludes this paper.

2 Task Description

This paper focuses on the task of textual entailment within the domain of clinical trial data analysis, specifically targeting Clinical Trial Reports (CTRs). CTRs serve as comprehensive documents containing essential information regarding various aspects of clinical trials, including eligibility criteria, interventions, results, and adverse events. Automating the analysis of CTRs through natural language processing techniques can significantly facilitate researchers' understanding and decision-making processes.

The task of NLI4CT involves analyzing annotated statements and determining their inference relation with the information contained in the CTR premises. These statements, characterized by an average length of 19.5 tokens, make claims about

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 432–436

various sections of the CTRs, including:

- **Eligibility criteria**: A set of conditions for patients to be allowed to take part in the clinical trial
- **Interventions**: Information concerning the type, dosage, frequency, and duration of treatments being studied.
- **Results**: Number of participants in the trial, outcome measures, units, and the results.
- **Adverse events**: These are signs and symptoms observed in patients during the clinical trial.

The NLI4CT task presents several challenges inherent to clinical trial data analysis, including numerical and quantitative reasoning, vocabulary and syntax variations, and comprehension of complex semantic structures. To address these challenges, interventions have been implemented targeting the following aspects:

- **Numerical Reasoning**: Models' abilities to apply numerical and quantitative reasoning are specifically targeted, given the importance of such inference in clinical trial analysis.
- **Vocabulary and Syntax**: Acronyms, aliases, and syntactic patterns common in clinical texts are addressed to improve model robustness and performance.
- **Semantics**: Complex reasoning tasks involving longer premise-hypothesis pairs are intervened upon to enhance model capabilities in handling intricate semantic structures.

3 System Description

To evaluate BERT on the SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials, we have fine-tuned BERT model on the NLI4CT dataset. we follow standard procedures for fine-tuning transformer-based models on natural language inference tasks. Here's a description of the process:

- **Data Preprocessing**: Tokenize the CTR premises and statements using the BERT tokenizer. Encode the tokenized sequences into input IDs, attention masks, and segment IDs as required by BERT.
- **Model Architecture**: Utilize the BERT architecture, which is a pre-trained transformer model. Add a classification layer on top of the BERT model to predict the entailment relation (entailment vs. contradiction) between the CTR premises and statements.
- **Fine-tuning Objective**: Fine-tune the pre-trained BERT model on the NLI4CT task using supervised learning. Minimize the binary cross-entropy loss between the predicted entailment labels and the ground truth labels.
- **Training Procedure**: Train the fine-tuned BERT model on the training data comprising CTR premises and statements along with their corresponding labels (entailment or contradiction).

4 Experimental Results

We experimented our model on on the SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. The experiment has been conducted in Google Colab environment[1], The following libraries: Transformers - Hugging Face[2] (Wolf et al., 2020), and Tensorflow[3] were used to train and to assess the performance of the model.

4.1 Datasets

The premises within NLI4CT are sourced from 1000 publicly accessible Breast cancer Clinical Trial Reports (CTRs) in English[4]. These records are overseen by the U.S. National Library of Medicine and adhere to the HIPAA Privacy Rule. The CTRs are categorized into four sections: Eligibility criteria, Intervention, Results, Adverse Events (Jullien et al., 2023).

A team of domain experts, including organizers of clinical trials from a prominent cancer research institution, participated in annotating the data. Annotators were tasked with generating entailment statements based on two CTR premises. These entailment statements make objectively true claims about the content of the premises. Annotators could choose to create statements regarding one or both premises. Substantial statements typically involve

[1] https://colab.research.google.com/
[2] https://huggingface.co/docs/transformers/index
[3] https://tensorflow.org
[4] https://ClinicalTrials.gov

summarization, comparison, negation, relation, inclusion, superlatives, aggregation, or rephrasing, requiring an understanding of multiple aspects of the premise. Annotators then select a subset of facts from the premises to support the claims in the statement.

Subsequently, a negative rewriting technique (Chen et al., 2020) was employed, altering the previously generated entailment statements to include objectively false claims while maintaining the original sentence structure and length. This technique aims to mitigate the likelihood of stylistic or linguistic patterns favoring either entailment or contradictory statements. Annotators then extract a subset of facts from the premises that contradict the claims in the false statement.

The resulting dataset comprises 2400 annotated statements with labels, premises, and evidence. The dataset was divided into train/test/dev sets in a 70/20/10 ratio. The two classes and four sections are evenly distributed across the dataset and its partitions (Jullien et al., 2023).

4.2 Evaluation Metric

The assessment of task performance will entail multiple stages. Initially, the performance on the original NLI4CT statements without any alterations, employing the Macro F1-score for evaluation.

Subsequently, the performance will be assess on the contrast set, comprising all statements with interventions. In this evaluation, two novel metrics—faithfulness and consistency—will be utilized, with their definitions provided below.

- **Faithfulness**: quantifies how accurately a system predicts outcomes for the right reasons. Essentially, it assesses the model's capacity to adjust its predictions accurately when encountering semantic-altering interventions. To compute faithfulness for a set of N statements x_i in the contrast set C, alongside their original statements y_i and model predictions f(), Equation 1 is utilized.

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

$$x_i \in C : Label(x_i) \neq Label(y_i), and f(y_i) = Label(y_i)$$

- **Consistency**: assesses how consistently a system generates identical outputs for problems that are semantically equivalent. Consequently, it gauges a system's capability to assign the same label to both original statements and contrast statements for interventions that preserve semantics. This means that even if the ultimate prediction is incorrect, the representation of the semantic phenomenon remains consistent across the statements. To calculate consistency for a set of N statements x_i in the contrast set C, along with their corresponding original statements y_i and model predictions f(), Equation 2 is employed.

$$Consistency = \frac{1}{N}\sum_{1}^{N} 1 - |f(y_i) - f(x_i)| \quad (2)$$

$$x_i \in C : Label(x_i) = Label(y_i)$$

4.3 Experimental Settings

During the fine-tuning of BERT model on the NLI4CT taining set, we set the hyper-parameters as follows: 10^{-5} as the learning rate, 30 epochs, 64 as the max sequence length, and 16 as batch size. Table 1 summarizes the hyperparameters settings of BERT base model.

Hyperparameters	Settings
Learning rate	10^{-5}
Batch size	16
Epochs	30
Max sequence length	64
Optimizer	Adam (Kingma and Ba, 2015)
Loss	Binary Cross Entropy

Table 1: Hyperparameters settings for the model in the experiments

4.4 System Performance

The reported results for the fine-tuned BERT model on the NLI4CT task are as follows:

- **Macro F1-score**: 0.62
- **Faithfulness**: 0.44
- **Consistency**: 0.54

The model achieved the 26th position in Macro F1-score, Faithfulness and Consistency among a total of 32 teams. The reported score of 0.62 in the Macro F1-score indicates that the model achieves moderate performance in accurately predicting the inference relation between CTR premises and statements. Moreover, the Faithfulness score, which is 0.44, suggests that the model struggles in making correct predictions for the right reasons. This indicates potential issues with reasoning or interpretation of the textual entailment task. On the other hand, the Consistency score, which is 0.54, indicates moderate consistency in the model's outputs for similar instances. However, there is room for improvement to achieve higher consistency.

The suboptimal performance of the fine-tuned BERT model on the NLI4CT task could be attributed to several factors: Firstly, clinical trial data, especially Clinical Trial Reports (CTRs), often contain domain-specific terminology, complex medical concepts, and nuanced language. BERT, being pre-trained on general-domain text, may struggle to comprehend and accurately reason over such specialized content. Secondly, The success of fine-tuning BERT depends on various hyperparameters, such as learning rate, batch size, and optimization algorithm. Suboptimal choices for these parameters can hinder convergence and degrade model performance. Thirdly, The interventions applied to the test set statements could introduce complexities or biases that the model is not equipped to handle, the model may struggle to generalize effectively. By addressing these factors the model performance can be improved in clinical trial data analysis tasks.

5 Conclusion

in this paper, an investigation is conducted into the utilization of BERT for NLI4CT, which underscores the complex nature of textual entailment tasks within the medical domain. The described approach tackles SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials. The model secured the 27th position among a total of 32 teams.

Despite challenges such as domain-specific terminology and nuanced semantics, our study reveals the potential for advancements in automated analysis of clinical trial reports. By recognizing the need for domain-specific approaches and leveraging the models, we pave the way for more accurate and reliable models tailored to the intricacies of medical data. Ultimately, our findings advocate for continued research and development efforts aimed at enhancing natural language processing techniques for clinical applications, thereby contributing to improved healthcare outcomes and medical decision-making processes.

References

Nancy Avis, Kevin Smith, Carol Link, Gabriel Hortobagyi, and Edgardo Rivera. 2006. Factors associated with participation in breast cancer clinical trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 24:1860–7.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jay DeYoung, Eric P. Lehman, Benjamin E. Nye, Iain James Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. *ArXiv*, abs/2005.04177.

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. 2023. 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Here's a breakdown of the results:

Control:

- **F1-score**: 0.6212
- **Recall**: 0.5899
- **Precision**: 0.6560

Contrast:

- **F1-score**: 0.4786
- **Recall**: 0.3655
- **Precision**: 0.6933

Faithfulness:

- **Score**: 0.4375

Consistency:

- **Score**: 0.5365

Paragraph Consistency:

- **Score**: 0.5813

Continuous Faithfulness:

- **Score**: 0.4160

Continuous Consistency:

- **Score**: 0.4080

Numerical Paragraph Consistency:

- **Score**: 0.5804

Numerical Continuous Faithfulness:

- **Score**: 0.5789

Numerical Continuous Consistency:

- **Score**: 0.6667

Definitions Consistency:

- **Score**: 0.5353

Paragraph:

- **F1-score**: 0.6293
- **Recall**: 0.5646
- **Precision**: 0.7107

Continuous:

- **F1-score**: 0.0
- **Recall**: 0.0
- **Precision**: 0.0

Numerical Paragraph:

- **F1-score**: 0.5
- **Recall**: 0.4845
- **Precision**: 0.5165

Numerical Continuous:

- **F1-score**: 0.0
- **Recall**: 0.0
- **Precision**: 0.0

Definitions:

- **F1-score**: 0.6001
- **Recall**: 0.5267
- **Precision**: 0.6973