SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials

Vaishnavi Asuri, Srija Vakiti
Department of Data Science, University of Colorado, Boulder, Colorado
{firstname.surname}@colorado,edu

Abstract

Large Language Models (LLMs) have shown exceptional performance in Natural Language Processing (NLP) tasks, but their susceptibility to shortcut learning, factual inconsistency, and performance degradation poses challenges, especially in critical domains like healthcare. This research focuses on advancing our understanding of LLM behavior in clinical settings, proposing a textual entailment task for Clinical Trial Reports (CTRs) to investigate robustness, consistency, and faithfulness in reasoning. The aim is to address the limitations of LLMs and enhance evaluation methodologies for clinical Natural Language Inference (NLI).

In response to the increasing deployment of LLMs in real-world scenarios, the study introduces a comprehensive analysis using controlled interventions, examining semantic phenomena in natural language and numerical inference. The research centers on the intersection of NLI and Clinical Trials (NLI4CT), emphasizing the importance of accurate interpretation and retrieval of medical evidence for personalized care.

The research conducted involves the organization of "SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data," which attracted substantial participation and submissions. However, the application of LLMs in critical domains demands further investigation and the development of robust evaluation methodologies.

This paper introduces the second iteration of NLI4CT, focusing on interventional and causal analyses of NLI models. The methodology enriches the dataset with a contrast set developed through interventions on the NLI4CT test set. Explicit causal relations between interventions and expected labels enable the exploration of model consistency and faithful reasoning in complex clinical NLI settings.

1 Introduction

Large language models (LLMs) have revolutionized natural language processing (NLP) but struggle with issues like shortcut learning, factual inconsistency, and vulnerability to distribution shifts. These challenges are especially critical in medical applications, where model errors can have severe consequences. This research addresses these concerns by proposing a textual entailment task using clinical trial reports (CTRs) to better understand LLM behavior in clinical settings.

Evaluating LLMs in real-world scenarios requires thoroughly investigating their robustness, consistency, and faithfulness in reasoning. By leveraging controlled interventions, the research systematically explores semantic phenomena in language and numerical inference. Studying natural language inference (NLI) on clinical trials is pivotal, offering a solution for large-scale interpretation and retrieval of medical evidence.

The paper presents the evolution of the NLI for clinical trials (NLI4CT) task, describing the organization of "SemEval-2023 Task 7" with extensive participation and submissions. While the first round produced high-performing LLM-based models, applying them in critical domains demands deeper understanding and more systematic analysis of model behavior and reasoning.

[1] https://codalab.lisn.upsaclay.fr/competitions/16190?secret_key=4863f655-9dd6-43f0-b710-f17cb67af607

[2] https://github.com/ai-systems/Task-2-SemEval-2024/tree/main

[3] https://sites.google.com/view/nli4ct/semeval-2024?authuser=0

The second round of NLI4CT introduces an intervention-based approach, augmenting the dataset using explicit causal relationships between interventions and expected labels. This facilitates exploring model consistency and faithful reasoning regarding clinical NLI. The paper outlines specific research aims and the task overview, emphasizing the necessity of advancing methodologies for evaluating LLMs in critical domains such as clinical trials.

2 Related Works

Numerous studies have delved into the challenges and capabilities of Large Language Models (LLMs) in the realm of Natural Language Processing (NLP), providing valuable insights into their strengths and limitations. Brown et al. (2020) and Chowdhery et al. (2022) established the state-of-the-art performance achieved by LLMs across various NLP tasks. However, investigations by Geirhos et al. (2020), Poliak et al. (2018), and Tsuchiya (2018) shed light on the pronounced susceptibility of LLMs to shortcut learning, raising concerns about their reliability in drawing accurate inferences. Furthermore, Elazar et al. (2021) highlighted issues related to factual inconsistency, while Miller et al. (2020) and Lee et al. (2020) identified performance degradation in the face of word distribution shifts and data transformations. These works collectively underscore the need for a deeper understanding of LLM behavior and the development of strategies to address these limitations.

In the specific context of medical applications, the work of Patel et al. (2008) and Recht et al. (2019) emphasized the potential overestimation of real-world performance by LLMs, making it imperative to scrutinize their reliability, especially in critical domains like healthcare. Recognizing the significance of robust evaluation methodologies, Wang et al. (2021) advocated for systematic behavioral and causal analyses to enhance the applicability of LLMs in real-world clinical trials. These foundational studies set the stage for the proposed research, which aims to advance our understanding of LLMs in the clinical setting through a focused textual entailment task.

The intersection of Natural Language Inference (NLI) and Clinical Trials (NLI4CT) has garnered attention, with seminal work by Bowman et al. (2015) presenting NLI as a solution for large-scale interpretation and retrieval of medical evidence. To address the practical challenges in the healthcare domain, Jullien et al. (2023) organized "SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data," attracting contributions from 40 participants and 23 participants in the entailment and evidence selection subtasks, respectively. The success of the first iteration, as evidenced by models developed by Zhou et al. (2023), Kanakarajan and Sankarasubbu (2023), and Vladika and Matthes (2023), achieving high performance with an F1 score of approximately 85%, underscores the potential of LLMs in NLI4CT. However, as outlined in the proposed research, the critical nature of real-world clinical trials necessitates further investigations and the development of evaluation methodologies grounded in systematic behavioral and causal analyses, echoing the sentiments put forth by Zhou et al. (2023).

3 Task Description

We are excited to take on this natural language inference task involving clinical trial reports in breast cancer research. As NLP researchers, evaluating model performance on complex medical texts is an important challenge. The multi-evidence structure using summarized CTR sections provides an intriguing setup to test textual entailment capabilities.

In approaching this task, we are particularly interested in the interventions applied to modify entailment relations in the test and development sets. Assessing how numerical reasoning, vocabulary, syntax, and semantic phenomena affect model judgments is key for developing robust clinical NLP methods. We plan to analyze these targeted interventions in depth to reveal when and why inconsistencies arise in reasoning across evidence documents.

This shared task focuses on multi-evidence natural language inference (NLI) using clinical trial reports (CTRs) from breast cancer studies. The CTRs are extracted from https://clinicaltrials.gov/ct2/home and annotated by domain experts.

1. Entailment and Evidence Selection Subtask:

This subtask involves determining the

inference relation (entailment vs. contradiction) between Clinical Trial Reports (CTR) and statements. The statements may make claims about a single CTR or compare two CTRs. The task is to determine the inference relation between CTR - statement pairs. The training set includes annotated statements that make claims about the information contained in different sections of the CTR premise.

2. Intervention Analysis Subtask:

This sub-task focuses on analyzing the robustness, consistency, and faithfulness of Natural Language Inference (NLI) models, particularly in the context of clinical NLI settings. The goal is to investigate the behavior of NLI models in their representation of semantic phenomena necessary for complex inference in clinical NLI settings.

For the sake of this project, we chose to focus on the former. The provided statements make claims about the information in one or more CTR sections. The average statement length is 19.5 tokens. Statements can refer to single or multiple CTRs to compare them. The task is to determine the inference relation (entailment or contradiction) between CTR-statement pairs.

Through participation in this shared task, we aim to produce advances in evidential NLP while elucidating model behavior on clinical trial data. We believe the interventions around numerical properties, vocabulary, syntax and semantics will drive progress on consistency, accuracy and interpretability for real-world deployment of language models.

4 Dataset

The dataset used for this research comprises statements and evidence generated by domain experts, clinical trial organizers, and research oncologists associated with the Cancer Research UK Manchester Institute and the Digital Experimental Cancer Medicine Team. Clinical Trial Reports (CTRs) form the core of the dataset, each consisting of four key sections: Eligibility criteria, Intervention details, Results, and Adverse events. These sections encompass vital information, including conditions for patient participation, treatment details, trial outcomes, and observed adverse events.

Within each CTR, the Inclusion and Exclusion criteria define the parameters for patient enrollment. For instance, the Inclusion Criteria encompass conditions such as the patient's age, type of breast cancer, previous treatments, and performance status. The Exclusion Criteria specify conditions that disqualify patients from participation, including specific cancer characteristics, medical history, and ongoing treatments. The dataset also includes detailed descriptions of two distinct interventions studied in a breast cancer clinical trial: Celecoxib and Placebo. Each intervention outlines dosage, frequency, and additional treatments, offering a comprehensive view of the trial design.

The Results section details the primary outcome measurements, focusing on Disease-Free Survival (DFS) as the primary endpoint for evaluating the interventions. This measurement encompasses loco-regional and distant breast cancer recurrence, new primary breast cancer occurrences, and death without disease relapse. Furthermore, the dataset provides specific results for each intervention arm, including the number of participants analyzed, 2-year DFS rates, and 5-year DFS rates.

The Adverse Events section presents a comprehensive overview of observed adverse events in both intervention arms. The dataset includes the total number of adverse events, as well as specific occurrences such as anemia, neutropenia, thrombocytopenia, and cardiac events. Each adverse event is documented with its frequency, providing valuable insights into the safety profile of the interventions.

5 Evaluation

The evaluation of model performance in this task involves a multifaceted approach to ensure a comprehensive understanding of the models' behavior. Initially, the assessment focuses on the original NLI4CT statements without any interventions, utilizing the Macro F1-score as the primary metric. This baseline evaluation provides insight into the inherent capabilities of the models in interpreting clinical trial data.

Subsequently, attention shifts to the contrast set, comprising statements subjected to various interventions. Two novel metrics, faithfulness and consistency, are introduced to assess model performance in this altered context. Faithfulness quantifies the extent to which a system produces correct predictions for the correct reasons when exposed to semantic-altering interventions. It is computed by comparing model predictions for the original and intervened statements in the contrast set, employing Equation 1. Consistency, on the other hand, measures the system's ability to produce consistent outputs for semantically equivalent problems, even if the final prediction is incorrect. This metric evaluates the uniformity of the model's representations across original and contrast statements for semantic-preserving interventions, calculated using Equation 2. The overall ranking of systems will be determined by averaging faithfulness and consistency scores across all intervention types.

The evaluation process consists of two phases: the 'practice' phase and the 'evaluation' phase. In the practice phase, participants' prediction files are assessed against the gold practice test set, offering an opportunity for refinement and calibration. Subsequently, during the evaluation phase, participants' prediction files are rigorously evaluated against the gold test set, reflecting real-world conditions. This dual-phase evaluation strategy ensures the robustness and generalizability of the models, considering their performance in both controlled practice scenarios and actual test conditions. Overall, the comprehensive evaluation framework aligns with the research objectives, emphasizing the need for models that not only excel in baseline performance but also demonstrate adaptability and reliability when confronted with varied and intervened clinical trial data.

6 Methodology

A. Traditional Models

The deployment of models in critical domains like clinical trials demands a careful choice of models that address specific linguistic and contextual nuances. In this approach, we leverage a combination of Word2Vec, TF-IDF, Countvectorizer, and SVM for a textual entailment task focused on Clinical Trial Reports (CTRs). The aim was to deploy relatively simpler models before reaching for more complex LLMs. This way, we can compare and contrast various models based on their outcomes and ease of implementation.

Word2Vec-based Logistic Regression

Model Description:

Word2Vec is chosen for its ability to capture semantic relationships between words in a continuous vector space.

Logistic Regression is employed as a classifier on top of Word2Vec embeddings.

Rationale:Word2Vec embeddings capture semantic meanings, essential for understanding the context of clinical trial language.

Logistic Regression is well-suited for binary classification tasks and complements the continuous embeddings.

TF-IDF-based Logistic Regression

Model Description:

TF-IDF (Term Frequency-Inverse Document Frequency) is utilized to represent the importance of words in a document.

Logistic Regression is applied to classify statements based on TF-IDF vectors.

Rationale:

TF-IDF accounts for the significance of terms in the corpus, providing a meaningful representation of document content.

Logistic Regression, being a linear model, is suitable for binary classification tasks and complements TF-IDF.

Countvectorizer and SVM

Model Description:

Countvectorizer converts text data into numerical vectors by counting the occurrences of words.

SVM (Support Vector Machine) is chosen as a classification model based on these numerical vectors.

Rationale:

Countvectorizer provides a straightforward representation of word occurrences, especially beneficial when exact word placement matters.

SVM is effective for high-dimensional data and binary classification, making it suitable for the task of textual entailment.

Model Integration and Prediction

Word2Vec-based Logistic Regression:

Utilized for capturing semantic relationships within the clinical trial language.

Logistic Regression helps in making binary predictions based on continuous embeddings.

TF-IDF-based Logistic Regression:

This was applied to consider the importance of words in clinical trial sections. Logistic Regression facilitates binary classification using TF-IDF vectors.

Countvectorizer and SVM:

Countvectorizer captures word occurrences, crucial for understanding the structure of clinical trial text. SVM is employed to classify statements into contradiction or entailment categories based on numerical vectors.

By employing a diverse set of models, we aim to account for the multifaceted nature of clinical trial language, capturing both semantic nuances and structural information. Each model is chosen with a specific purpose, contributing to a comprehensive understanding of the textual entailment task within the domain of Clinical Trial Reports.

B. Approach B: Deep Learning BERT models

Our primary goal, through this approach, was to develop a robust model capable of discerning between different types of statements, classifying them into two fundamental categories: "Contradiction" and "Entailment." This classification not only aids in deciphering nuanced biomedical content but also lays the groundwork for automating the comprehension of complex textual information.

i. Data Preprocessing: First, the biomedical text data, consisting of statements and corresponding labels (Contradiction/Entailment), was loaded from JSON files for both the training and development sets. Instances were formatted into a suitable structure for training and evaluation.

ii. Model Selection: The BERT (Bidirectional Encoder Representations from Transformers) model was chosen for its effectiveness in capturing contextual information in text. The 'bert-base-uncased' pre-trained model and tokenizer were utilized.

iii. Dataset Creation: A custom dataset class, CustomDataset, was implemented to tokenize and encode text statements using the BERT tokenizer. The dataset class included functionalities for handling input statements, labels, and padding to a specified maximum length.

iv. Model Configuration: The BERT model for sequence classification (Bert For Sequence Classification) was used, assuming a binary classification task (Contradiction/Entailment). The model was initialized with the appropriate number of labels and loaded with the pre-trained weights.
v. Training Loop:The training loop involved iterating over batches from the training dataset using a DataLoader. For each batch, the BERT model received input_ids, attention_mask, and labels, performed a forward pass, calculated loss, and executed backward pass for optimization. The AdamW optimizer was employed for parameter updates.
vi. Evaluation Loop: The evaluation loop was similar to the training loop but excluded the backward pass. Accuracy was computed for each batch during the evaluation on the development set.
vii. Evaluation Metrics: Beyond accuracy, additional evaluation metrics, including F1 score, precision, and recall, were calculated on the entire development set. These metrics provide a comprehensive understanding of the model's performance in terms of precision and recall for each class.
viii. Results: The final evaluation metrics (F1, precision, and recall) were reported, offering insights into the model's ability to classify biomedical text statements as either Contradiction or Entailment.

7 Results

A. Approach A

The analysis of the results from the Word2Vec (W2V) based Logistic Regression and TF-IDF based Logistic Regression models provides valuable insights into their performance in the textual entailment task focused on Clinical Trial Reports (CTRs).

Word2Vec-based Logistic Regression:
The Word2Vec model with Logistic Regression achieved an accuracy of 48.24%. The classification report reveals a balanced precision and recall for both contradiction and entailment categories. Notably, the model exhibits higher recall (71%) for entailment compared to contradiction (26%), suggesting a better ability to capture true entailments. However, the overall F1-score indicates a moderate performance (0.45), and the macro and weighted averages hover around 0.46 and 0.48, respectively. The relatively low F1-score emphasizes the need for further refinement in capturing both precision and recall.

TF-IDF based Logistic Regression:
In contrast, the TF-IDF based Logistic Regression model achieved a lower accuracy of 35.88%. Similar to the Word2Vec model, the classification report indicates a challenge in achieving a balance between precision and recall. Both contradiction and entailment categories show comparable precision and recall values. The F1-score, macro, and weighted averages around 0.36 reflect a modest overall performance. The lower accuracy and F1-score suggest that the TF-IDF based approach may encounter difficulties in effectively capturing the nuances of textual entailment in the context of clinical trial reports.

Comparative Analysis:
Comparing the two models, the Word2Vec based Logistic Regression model outperforms the TF-IDF based counterpart in terms of accuracy. However, both models face challenges in achieving a harmonious balance between precision and recall, as indicated by the F1-scores. The choice between these models should be guided by the specific requirements of the application, considering factors such as the importance of true positives, false positives, and false negatives in the context of clinical trial textual entailment. Further optimization and exploration of model parameters may contribute to enhanced performance and robustness in capturing subtle linguistic relationships within clinical trial reports.

Test Set:
In the final evaluation of the test set, the model exhibited a modest F1 score of 0.227273, reflecting a balanced performance between precision and recall. The precision score was measured at 0.468750, indicating the model's ability to correctly identify positive instances among its predictions. However, the recall score was comparatively lower at 0.150000, suggesting that the model missed a notable proportion of actual positive instances. The overall accuracy achieved on the test set was 0.490000, indicating the ratio of correctly predicted instances to the total number of instances. These metrics collectively provide insights into the model's performance, highlighting its strengths and areas for improvement in handling the specific characteristics of the test data.

B. Approach B

The observed validation accuracies during the model evaluation on the biomedical dataset varied across different batches. The recorded values ranged from 0.25 to 0.875, showcasing fluctuations in the model's performance on distinct subsets of the validation data. Instances of lower accuracy, such as 0.25, suggest challenges or complexities in correctly classifying certain statements within the dataset. Conversely, instances of higher accuracy, such as 0.875, indicate the model's adeptness in accurately predicting the labels for specific batches.

In the retrospective evaluation of the BERT-based model on the final set, the model demonstrated a promising F1 score of 0.67, indicating its proficiency in achieving a harmonious balance between precision and recall. The precision score, a metric measuring the accuracy of positive predictions, was reported at 0.53, suggesting the model's capability to make accurate assertions in identifying entailment instances. Furthermore, the recall score, representing the model's sensitivity in capturing relevant positive instances, achieved an impressive value of 0.90, showcasing the model's effectiveness in detecting a substantial proportion of actual entailment instances.

Test Set:
In the comprehensive evaluation of the test set, the model demonstrated an F1 score of 0.666667, reflecting a harmonious balance between precision and recall. The precision score, measured at 0.528302, indicated the model's capability to accurately identify positive instances among its predictions. The recall score, at 0.903226, signified the model's effectiveness in capturing a substantial proportion of actual positive instances. Despite these strengths, the overall accuracy achieved on the test set was 0.500000, indicating the ratio of correctly predicted instances to the total number of instances. These evaluation metrics provided valuable insights into the model's performance on the specific characteristics of the test data, demonstrating its proficiency in certain aspects while pointing towards potential areas for refinement.

8 Limitations

The traditional models like Word2Vec logistic regression and TF-IDF logistic regression achieved only modest performance in terms of accuracy and F1 scores on the clinical trial textual entailment task. This indicates inherent challenges in capturing the complex semantic and syntactic relationships within clinical trial reports using these conventional approaches.

Both the Word2Vec and TF-IDF models

struggled to balance precision and recall, making trade-offs between the two. This highlights difficulties in simultaneously maximizing true positives and minimizing false positives for textual entailment prediction in the clinical domain.

The BERT model demonstrated significant fluctuations in accuracy across validation batches. The variability in performance across subsets of data points to overfitting particular patterns. This poses generalizability issues when exposed to previously unseen statements.

Despite promising validation results, the BERT model exhibited a noticeable dip in recall score when evaluated on the test set. The sharp decline shows the model's brittleness and suggests it may latch onto spurious cues during training that fail to transfer.

All the developed models have only been tested on a small dataset of clinical trial reports focusing on breast cancer studies. Their real-world effectiveness across diverse medical domains remains unverified.

9 Conclusion

While the initial results are promising, this research emphasizes that considerable work remains to enhance LLM robustness for deployment in critical real-world applications. As evident from the limitations discussed, reliance on spurious patterns can restrict generalization capability. Moving forward, the focus must shift towards models that can capture the nuances of clinical language, perform sound quantitative reasoning, and make consistent judgments in the face of complex, multi-hop inferences.

By releasing the datasets, findings and insights from this study, it is hoped that the NLP community will build on these efforts to produce reliable, transparent and practical LLMs for high-stakes medical settings. Such models can pave the path towards personalized evidence-based care by facilitating accurate interpretation and retrieval of clinical trial knowledge at scale.

References

[1] Wallace, Byron C.. “What Does the Evidence Say? Models to Help Make Sense of the Biomedical Literature.” IJCAI : proceedings of the conference 2019 (2019): 6416-6420 .

[2] Stammbach, Dominik. “Evidence Selection as a Token-Level Prediction Task.” Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER) (2021): n. pag.

[3] Müller, Martin, Marcel Salathé and Per Egil Kummervold. “COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter.” Frontiers in Artificial Intelligence 6 (2020): n. pag.

[4] Jin, Qiao, Chuanqi Tan, Mosha Chen, Xiaozhong Liu and Songfang Huang. “Predicting Clinical Trial Results by Implicit Evidence Integration.” ArXiv abs/2010.05639 (2020): n. pag.

[5] DeYoung, Jay, Eric P. Lehman, Benjamin E. Nye, Iain James Marshall and Byron C. Wallace. “Evidence Inference 2.0: More Data, Better Models.” ArXiv abs/2005.04177 (2020): n. pag.

[6] Kanakarajan, Kamal Raj, Bhuvana Kundumani and Malaikannan Sankarasubbu. “BioELECTRA:Pretrained Biomedical text Encoder using Discriminators.” Workshop on Biomedical Natural Language Processing (2021).

[7] Houssein, Essam H., Rehab E. Mohamed, and Abdelmgeid A. Ali. "Machine learning techniques for biomedical natural language processing: a comprehensive review." IEEE Access 9 (2021): 140628-140653.

[8] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So and Jaewoo Kang. “Data and text mining BioBERT: pre-trained biomedical language representation model for biomedical text mining.” (2019).

[9] Beltagy, Iz, Kyle Lo and Arman Cohan. “SciBERT: A Pretrained Language Model for Scientific Text.” Conference on Empirical Methods in Natural Language Processing (2019).

[10] Chiu, Billy, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. "How to train good word embeddings for biomedical NLP." In Proceedings of the 15th workshop on biomedical natural language processing, pp. 166-174. 2016.