Stanford MLab at SemEval 2023 Task 7: Neural Methods for Clinical Trial Report NLI

Conner Takehana, Dylan Lim, Emirhan Kurtulus, Ramya Iyer, Ellie Tanimura, Pankhuri Aggarwal, Molly Cantillon, Alfred Yu, Sarosh Khan, Nathan Chi*, and **Ryan A. Chi***
Stanford University
{conner7, dylanlim, emirhank, ramya.iyer, etanim, pankhuri, mcan, ayu1001, skhan44, nchi1, ryanchi} @stanford.edu

Abstract

We present a system for natural language inference in breast cancer clinical trial reports, as framed by SemEval 2023 Task 7: *Multi-evidence Natural Language Inference for Clinical Trial Data*. In particular, we propose a suite of techniques for two related inference subtasks: entailment and evidence retrieval. The purpose of the textual entailment identification subtask is to determine the inference relation (either entailment or contradiction) between given statement pairs, while the goal of the evidence retrieval task is to identify a set of sentences that support this inference relation. To this end, we propose fine-tuning Bio+Clinical BERT, a BERT-based model pre-trained on clinical data. Along with presenting our system, we analyze our architectural decisions in the context of our model's accuracy and conduct an error analysis. Overall, our system ranked 20 / 40 on the entailment subtask.

1 Introduction

Clinical trial reports (CTRs) detail the treatments and reactions that patients undergo over the course of a clinical study. The growing corpus of CTRs, coupled with a lack of adequate analytical tools, makes it increasingly difficult for clinical practitioners to provide up-to-date, personalized, evidence-based care (DeYoung et al., 2020). Improvements in the reliability of CTRs have not been accompanied by developments in the ability to analyze and compare them, leading to a substantial disparity between the quality of CTRs and physicians' abilities to capitalize on them. (Butcher et al., 2019).

While natural language inference (NLI) models demonstrate impressive capabilities in inferring logical relationships between texts, applications in the medical domain come with a unique set of challenges (Percha et al., 2021). One issue is that accurately inferring relationships between entities in clinical data requires knowledge of domain-specific jargon (Sushil et al., 2021). Additionally, because scientific communication can be highly variable in its technical language, abbreviations, and formatting, successful NLI systems must be trained on a large representative corpus of varied data. Finally, it can be difficult for NLI systems to distinguish between non-entailed subsequences like "John believes Martha is sick" and "John believes Martha" (McCoy and Linzen, 2018).

The precise interpretation and extraction of evidence from CTRs could drastically diminish the current disconnect between research findings and clinical practices, as it will be much easier for clinicians to parse new publications and stay updated on the most effective treatment practices (**?**). In this paper, we experiment with a variety of modern and traditional machine learning approaches to accomplish this task. Ultimately, we find that Bio+Clinical BERT and BioBERT achieve the highest performance, obtaining an F1 score of 0.662 on the evaluation set.

2 Background

2.1 Task Overview

Data for this task is made available through SemEval Task 7 (Jullien et al., 2023). The 2400 statements for the dataset are collected by clinical trial organizers, domain experts, and research oncologists from the Cancer Research UK Manchester Institute and the Digital Experimental Cancer Medicine Team.

Each CTR contains four main sections:

- **Eligibility Criteria** describes the required conditions of patients in the clinical trial.
- **Intervention** explains the treatment type, quantity, frequency, and duration used in the trial.

*Co-senior authors.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1769–1775

- **Results** reports the number of participants and outcome details including measures, units, and results.

- **Adverse Events** outlines the signs and symptoms observed during the course of the trial. For Task 1, the dataset consists of a pair of CTRs, a statement, a section marker, and an entailment/contradiction label.

2.2 Prior Work

Previous work considered using NLI for the adjacent task of clinical registry curation tested the performance of five models, the highest performing of which was a BERT-based model (Percha et al., 2021). Given the limited size of training corpora, BERT-based transfer learning approaches are generally very well suited for the biomedical domain (Kanakarajan et al., 2021).

For biomedical text-related tasks, the domain-specific fine-tuned Bio BERT model outperforms BERT and previous state-of-the-art models (Lee et al., 2020). On small training sets, Med-BERT has been shown to improve prediction accuracy (Rasmy et al., 2021). For text in clinical trials, a domain-specific language model PubMedBERT outperformed BERT and its variants for NER but the difference amongst models was small demonstrating that domain-specific BERT trained on clinical breast radiology reports resulted in increased performance in NLP text sequence classification tasks over generic BERT embeddings such as classic BERT fine-tuned to the same tasks (Li et al., 2022; Kuling et al., 2022). With CancerBERT, an oncological domain-specific language model, Zhou et al. (2022) demonstrated that increased granularity and specificity in the training set further improves the performance of domain-specific BERT models in clinical NLP tasks (Zhou et al., 2022).

Gu found that domain-specific pre-training from scratch, as opposed to BERT models that generate their vocabulary only partially from biomedical text, yielded more accuracy in evidence-based medical information extraction and relation extraction NLP tasks (Gu et al., 2021). At various natural language tasks like NLI and relation extraction in the biomedical domain, BioELECTRA, a fine-tuned ELECTRA-based model, outperformed BERT and ALBERT (Kanakarajan et al., 2021).

3 System Overview

3.1 Task 1: Learning Models

- **Bio+Clinical BERT** (Alsentzer et al., 2019) is a BERT-based model trained on all notes and articles from the MIMIC-III clinical database (roughly 880M words). The model uses a fine-grained tokenization approach that is able to identify and separate subword units, which is particularly useful for handling long and complex medical terms. The model was trained using two pre-training tasks: masked language modeling and next-sentence prediction.

- **RoBERTa** (Liu et al., 2019) is a BERT-based transformers model, that was pre-trained using masked language modeling and dynamic masking. The model randomly masks 15% of the words in a sentence, then processes the entire masked sentence and predicts the missing words. This enables the model to hold a bidirectional representation of the sentence.

- **DistilBERT Base Uncased Finetuned SST-2 English** (Sanh et al., 2019) is a BERT-based model that was fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset for sentiment analysis.

- **ELECTRA Small Discriminator** (Clark et al., 2020) is a model that utilizes discriminative pre-training to improve efficiency and reduce computational resources. The model holds a small size and number of parameters in comparison to other pre-trained language models, making it more computationally efficient and faster to fine-tune downstream tasks.

- **BioBERT v1.1** (Lee et al., 2020) is a pre-trained language model based on BERT Architecture, fine-tuned on various biomedical tasks, including named entity recognition, relation extraction, and biomedical question answering.

- **BioMed NLP PubMedBERT Base Uncased** (Gu et al., 2021) is a pre-trained language model that seeks to improve the discrepancies in BERT by pre-training the model from scratch results rather than general domain corpora.

- **Bio Discharge Summary BERT** (Alsentzer et al., 2019) is a BERT-based model which

was specifically trained on discharge summaries from MIMIC-III. The model is fine-tuned for various clinical natural language processing tasks, such as named entity recognition and relation extraction.

- **BioBERT Diseases NER** (Lee et al., 2020) is a BERT-based pre-trained named entity recognition model designed to identify disease entities. The model was fine-tuned in the NER task with BC5CDR-diseases and NCBI-diseases corpus.
- **MedBERT Breast Cancer** (Rasmy et al., 2021) is a pre-trained language model specifically designed for processing clinical notes related to breast cancer.
- **GloVe** (Pennington et al., 2014) is a model for distributed word representations. We use bag-of-wards GloVe embeddings as input to our random forest and support vector classifiers.

3.2 Task 1: Data Augmentation

As a small quantity of training data was given (1700 CTRs for test and 200 for evaluation), data augmentation is performed using the methods detailed below.

3.2.1 Back Translation

We apply back translation to generate new data. This method involves translating data to a foreign language and then back to the starting language. This allows for new CTRs to be created with different syntactic and semantic structures. By increasing the quantity of data and diversity of languages, the model can better generalize to unseen test data.

Using Google Translate, we back translate and add 1700 additional CTR data points to our training set. All 1700 data points were used to train the model.

3.2.2 Synonym Replacement

We also explore using synonym replacement as a form of data augmentation. Synonym replacement creates new data points by replacing words or phrases in a text with their synonyms while preserving the overall meaning of the original text.

We perform synonym replacement using NLTK (Bird and Loper, 2004). Through NLTK, WordNet, a large lexical database for English, is used to generate lists of synonyms for given words in CTRs. To create new data points, words in CTRs are replaced with synonyms with high similarity. These synonyms are chosen through pre-trained word embeddings from BioWordVec and BioSentVec. Word embeddings are vectors in high-dimensional space that capture the semantic relationships between words based on their co-occurrence patterns in text. BioWordVec and BioSentVec are trained by PubMed articles and clinical notes from the MIMIC-III Clinical Database. We create 1700 additional CTRs through synonym replacement.

3.2.3 Random Perturbation

Random insertions, deletions, and swapping of words are also performed. Random insertions add noise to help models better understand relevant points of the text. Random deletions can prevent overfitting on specific words and phrases. Random swaps allow models to learn different syntactical arrangements without corrupting the meaning of the CTR. We apply all three augmentations to our training set with custom-made algorithms.

Insertions are performed by dividing the text into 75-word segments, randomly choosing phrases that have a maximum of two words in each segment, and then randomly inserting them back into the segment. Deletions are randomly applied at intervals of 1 deletion for every 25 words. Swapping is done by randomly selecting and swapping 2 words for every 30 words. We create 1700 additional datapoints through random insertions, deletions, and swapping.

3.3 Task 1: Hyperparameters

3.3.1 Optimizers

We use two optimizers for our models: cross-entropy loss and Sharpness-Aware Minimization (SAM). We use cross-entropy loss beacuse it is the standard optimizer for classification problems. The Sharpness-Aware Minimization (SAM) optimizer takes into account the sharpness of the loss function during optimization to find flatter minima. We test the SAM optimizer sharpness-awareness metric, `sam_rho`, for values ranging from 0.01 to 0.05.

3.3.2 Learning Rates

Given that many models during pilot testing only exhibited random guessing, we explore a variety of learning rates. We assume that random guessing is the result of large learning rates overshooting minima or small learning rates failing to converge. Therefore, we test learning rates ranging from 10^{-2} to 10^{-6}.

Data Augmentation	Original Text	Augmented Text
Back Translation	The human epidermal growth factor receptor 2 (HER2) status of the tumor will be used to stratify patients.	Human epidermal growth factor receptor 2 (HER2) tumor status will be used to stratify patients.
Synonym Replacement	Patients must have histologically or cytologically confirmed breast cancer with metastatic disease.	Patients must have microscopically or cytologically confirmed breast cancer with disseminated disease.
Random Perturbation	Prior Therapy: Patients must not have received prior chemotherapy in the metastatic breast setting.	Prior Therapy: Patients must not have received prior received in the metastatic chemotherapy breast setting.

Table 1: Examples of the three data augmentations applied to different sentences.

3.4 Task 1: Ensembling

After all individual models are tested, we train the top five highest-performing models twice, each on different seeds. We use the 10 subsequent models to create a single model through voting ensembling. We perform voting ensembling by taking the outputs of all 10 models and choosing the output with the most 'votes.' This adjusts for models that consistently underperform on specific classification inferences, which other models could make up for.

3.5 Task 2: Evidence Retrieval

To justify the label predicted in Task 1, we aimed to extract the supporting facts from a CTR premise to substantiate the decision and analyze the effectiveness of the BM25 evidence retrieval baseline for the task.

To achieve this goal, we used a dataset consisting of CTR premises and corresponding statements along with manually annotated supporting facts. We implemented the BM25Okapi algorithm to retrieve the relevant evidence from the primary and secondary sections of the CTR premise. We set a threshold of 1 to retain only the evidence with BM25 scores above this value.

The dataset was randomly split into training and development sets, and we trained the BM25 model on the training set. We then evaluated the model on the development set using standard evaluation metrics – precision, recall, and F1 score.

4 Experimental Setup

4.1 Task 1

For the subtask, we utilized the provided train and validation split in our experimental setup. Other than tokenization, we did not perform substantial preprocessing on the data. GloVe, in conjunction with, traditional ML methods (random forest classifier and support vector classifier) is used as a baseline for other models.

During the phase of tuning hyperparameters, we assessed models by means of the token-level F1 score, which we computed using PyTorch's metrics module to generate a classification report. As a result of the iteration constraint, hyperparameters were tuned by hand. We found main success through the implementation of varying optimizers and learning rates.

Two optimizers were used for our models: The cross-entropy loss and Sharpness-Aware Minimization (SAM) optimizer. Cross-entropy loss was used as it is the standard optimizer for classification problems. The Sharpness-Aware Minimization (SAM) optimizer takes into account the sharpness of the loss function during optimization to find flatter minima. The SAM optimizer sharpness-awareness metric, sam_rho, was tested for values ranging from 0.01 to 0.05.

Learning rates were a large subject of conversation due to many models' classification never learning beyond random guessing. Therefore, learning rates ranging from 10^{-2} to 10^{-6} were tested. We found that the learning rate of 10^{-4} was most successful.

4.2 Task 2

The data is first preprocessed into its components: statement, type, primary evidence id, section id, and optionally a secondary evidence id. Then, we load the `BERTForSequenceClassification`

model and train it on a large corpus of text. The tokenizer used was `BERTTokenizer`, which helped fine tune the model on the training data.

We first retrieve the primary and secondary evidence from the JSON files using the provided ids. If the trial type is "Comparison", then secondary evidence is also retrieved. The statement, primary evidence, and secondary evidence (if present) are concatenated into one string. The combined text is then tokenized using the `BERTTokenizer` and encoded using the `encode_plus` function. The encoded `input_ids` and `attention_masks` are then passed through the `BERTForSequenceClassification` model to obtain the logits.

The cosine similarity between the logits of the statement and the logits of the concatenated text is then computed using the functional module in PyTorch. If the cosine similarity is greater than a pre-defined threshold, the trial is considered to have a match. The threshold used in this experiment is set to 0.80.

The results are then stored in a dictionary where the key is the index of the trial in the dev set and the value is the cosine similarity between the statement and the combined text. The trials that have a cosine similarity greater than the threshold are then filtered and stored in a list of tuples. Each tuple contains the index of the trial and its cosine similarity.

5 Results

5.1 Task 1: Individual Model Performance

From table 2, we see that the top 5 performing models are PubMedBert (F1 = 0.660), RoBERTa Base (F1 = 0.658), BioBERT Disease (F1 = 0.648), Bio+Clinical BERT (F1 = 0.648), BioBERT Disease (F1 = 0.655), and ELECTRA Small (.639).

We find that BERT-based models such as PubMedBERT and Bio+Clinical BERT have the strongest performance. Additionally, the models trained on medical data generally outperform other models. We also find that Cross-entropy loss outperforms the SAM optimizer by 11.5 accuracy percentage points. Overall, the SAM optimizer was not able to outperform random classification.

Furthermore, we see that all best-performing models have learning rates of 10^{-4}. However, when models fail to learn, smaller and bigger learning rates allow for non-random classification. For example, PubMedBERT is not able to learn with a learning rate of 10^{-4}. But with a learning rate of 10^{-6}, an F1 score of 0.503 and accuracy of 59.5% are achieved.

5.2 Task 1: Voting Ensembling

The 5 models listed in 5.1 trained twice each and undergo voting ensembling. The voting ensembling produces the final outputs for submission. This ensembling produces an F1 score of 0.662, higher than any of the F1 scores from individual models. This placed us as the 20th out of 30 submissions to Task 1.

5.3 Task 2: Evidence Retrieval Evaluation

Our model obtains the results shown below, from the evaluation of the BM25 evidence retrieval baseline on the development set.

These results demonstrate that the BM25 evidence retrieval baseline is somewhat effective in retrieving relevant evidence from the CTR premise to justify the label predicted in Task 1. The precision and recall scores suggest that the model is able to retrieve the majority of the relevant evidence while avoiding irrelevant information. However, there is potential for model performance to improve, in terms of the F1 score.

6 Conclusion

In this paper, we evaluate a variety of BERT-based and other models for two inference sub-tasks: entailment and evidence retrieval. The models are compared on their accuracies and F1 scores. Clinical trial reports from SemEval Task 7 dataset are used for the task. Due to the small training dataset size, we use back translation, synonym replacement, and random perturbation to augment our training data.

The top 5 performing models for Task 1 are Bio+Clinical BERT, DistilBERT, BioBERT, ELECTRA Small, and Bio Discharge Summary BERT. Each of these has a learning rate of 10^{-4}. These models are used to create a single voting ensemble model. The ensembling model results in an F1 score of 0.662 and outperforms other models. For the BM25 evidence retrieval baseline, our model is able to retrieve the majority of the relevant evidence. We use the voting ensembling model to make our final predictions.

7 Concerns, Limitations, and Extensions

The data augmentation performed in this paper combined synthesized augmented clinical trial data

Model	Learning Rate	F1 Score	Accuracy
Bio+Clinical BERT	10^{-2}	.601	50.0%
Bio+Clinical BERT	10^{-4}	.648	62.0%
Bio+Clinical BERT	10^{-6}	.532	53.5%
RoBERTa Base	10^{-4}	.648	50%
RoBERTa Base	10^{-6}	.658	50%
DistilBERT Base	10^{-4}	.608	**64.5**%
DistilBERT Base	10^{-6}	.570	55.5%
ELECTRA Small	10^{-4}	.639	62.5%
ELECTRA Small	10^{-6}	.592	52.5%
BioBERT v1.1	10^{-4}	.637	59.5%
BioBERT v1.1	10^{-6}	.592	52.5%
PubMedBERT	10^{-4}	**.660**	50.0%
PubMedBERT	10^{-6}	.503	59.5%
Bio Discharge BERT	10^{-4}	.592	62.5%
Bio Discharge BERT	10^{-6}	.359	53.5%
Biomedical NER	10^{-4}	.610	58.5%
BioBERT Diseases NER	10^{-4}	.655	61.0%
BioBERT Diseases NER	10^{-6}	.648	54.0%
MedBERT Breast Cancer	10^{-4}	.314	56.5%
GloVe + Random Forest	–	.601	57.5%
GloVe + Support Vector	–	.628	58.5%

Table 2: Model performances based on their learning rates, evaluated by F1 and normal accuracy percentages.

Model	F1 Score	Precision Score	Recall Score
Ensembled models	.662	.575	.780

Table 3: The ensembled model is created by training the top 5 performing models outlined in 5.1 twice, and performing voting ensembling.

Optimizer	sam_rho	Accuracy
Cross-entropy loss	-	62.0%
SAM	0.01	49.0%
SAM	0.03	50.5%
SAM	0.05	50.0%

Table 4: Comparison of Cross-entropy loss and SAM optimizer used on the Bio+Clinical BERT model.

Metric	Score
Precision	0.522748
Recall	0.622034
F1	0.461280

Table 5: Model performance on Task 2, evaluated by F1, precision and recall.

and combined it with the originally provided clinical data. As no strict regulation was performed on screening the augmented data, the question is brought up of its accuracy. Merging the two datasets together and utilizing them as one may have led to the development of skewed results.

For Task 1, we were only able to perform one run with a 10^{-2} learning rate. Furthermore, we only tested the SAM optimizer with our Bio+Clinical BERT model. Further testing of these hyperparameters would be appropriate. In addition, another possible extension would be to generate word embeddings with the BioWordVec machine learning model and compare it to the accuracy achieved by the GloVe model.

There are no ethical concerns with out work.

8 Acknowledgements

This research effort would not have been possible without the support of Stanford ACM. The authors thank Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and Andre Freitas for organizing SemEval Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Nancy J Butcher, Andrea Monsour, Emma J Mew, Peter Szatmari, Agostino Pierro, Lauren E Kelly, Mufiza Farid-Kapadia, Alyssandra Chee-A-Tow, Leena Saeed, Suneeta Monga, et al. 2019. Improving outcome reporting in clinical trial reports and protocols: study protocol for the instrument for reporting planned endpoints in clinical trials (inspect). *Trials*, 20:1–10.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Mael 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*.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Grey Kuling, Belinda Curpen, and Anne L Martel. 2022. Bi-rads bert and using section segmentation to understand radiology reports. *Journal of Imaging*, 8(5):131.

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.

Jianfu Li, Qiang Wei, Omid Ghiasvand, Miao Chen, Victor Lobanov, Chunhua Weng, and Hua Xu. 2022. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Medical Informatics and Decision Making*, 22(S3).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

R. Thomas McCoy and Tal Linzen. 2018. Non-entailed subsequences as a challenge for natural language inference.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bethany Percha, Kereeti Pisapati, Cynthia Gao, and Hank Schmidt. 2021. Natural language inference for clinical registry curation. *medRxiv*, pages 2021–06.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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*, pages 41–53.

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. 2022. Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216.