Clemson NLP at SemEval-2023 Task 7: Applying GatorTron to Multi-Evidence Clinical NLI

Ahmed Alameldin*
School of Computing
Clemson University
aalamel@clemson.edu

Ashton Williamson*
School of Computing
Clemson University
taw2@clemson.edu

Abstract

This paper presents our system descriptions for SemEval 2023-Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data sub-tasks one and two. Provided with a collection of Clinical Trial Reports (CTRs) and corresponding expert-annotated claim statements, sub-task one involves determining an inferential relationship between the statement and CTR premise: contradiction or entailment. Subtask two involves retrieving evidence from the CTR which is necessary to determine the entailment in sub-task one. For sub-task two we employ a recent transformer-based language model pretrained on biomedical literature, which we domain-adapt on a set of clinical trial reports. For sub-task one, we take an ensemble approach in which we leverage the evidence retrieval model from sub-task two to extract relevant sections, which are then passed to a second model of equivalent architecture to determine entailment. Our system achieves a ranking of seventh on sub-task one with an F1-score of 0.705 and sixth on sub-task two with an F1-score of 0.806. In addition, we find that the high rate of success of language models on this dataset may be partially attributable to the existence of annotation artifacts. [1]

1 Introduction

The proliferation of Clinical Trial Reports (CTRs) has made it challenging for medical practitioners to stay up-to-date on the latest literature, inhibiting their ability provide evidence-based patient care. The computerized clinical decision support system (CDSS) has aimed to address this problem by providing an interface, typically within the electronic health record system, with which a clinical practitioner can, for example, access automated patient care recommendations or view diagnostic suggestions (Sutton et al., 2020). Research suggests that these systems can have a positive impact on both clinical practitioners and patients (Kwan et al., 2020; Sutton et al., 2020); they depend, however, on the ability to interpret and retrieve information from biomedical literature (Sutton et al., 2020). In this context, Task 7 provides a chance to augment existing approaches for retrieval and interpretation of unstructured biomedical text.

Task 7 is split into two sub-tasks which are structured around a dataset of English-language breast cancer CTRs: *sub-task 1: textual entailment* and *sub-task 2: evidence retrieval*. Sub-task one involves identifying the entailment relation (entailed or contradicted) between a statement–a claim about the CTR(s) annotated by a domain-expert–and a premise in the form of a section from the CTR(s). Sub-task 2 involves extracting sections of text from the CTR section which are necessary to either entail or contradict the statement. These tasks prove challenging by requiring modeling of long-term dependencies, comparisons between multiple CTRs, and complex numerical reasoning.

In this paper, we outline our approach to the Multi-evidence Natural Language Inference for Clinical Trial Data shared task and examine our results. For both sub-tasks, our system uses GatorTron, a recently introduced transformer-based clinical language model, as a foundation model due its demonstrated success at medical natural language inference tasks (Yang et al., 2022) as well as to explore the impact of increased model parameters on a down-stream clinical task. To overcome the distribution shift between the pretraining domain and the target domain, we domain adapt the model on a collection of clinical trial reports similar to the task dataset (Xu et al., 2021). For task 1–textual entailment–we improve base performance by restricting the CTRs to text necessary to determining entailment, as identified by our model for sub-task 2. For sub-task 2, we framed the problem as a binary classification task by splitting each

[1] Pretrained model weights are available at https://huggingface.co/AshtonIsNotHere

[*] These authors contributed equally to this work.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1598–1602
July 13-14, 2023 ©2023 Association for Computational Linguistics

CTR premise into individual lines. Our systems placed seventh on sub-task one with an F1-score of 0.705 and sixth on sub-task two with an F1-score of 0.806. Additionally, we explore the existence of annotation artifacts (Gururangan et al., 2018) in the dataset, finding that our system for sub-task 1 performs well above the baseline when predicting on statements only (F1-score 0.584).

2 Dataset

The organizers of Task 7 provide a dataset of 1,000 English-language, breast cancer clinical trial reports extracted from https://clinicaltrials.gov/ct2/home along with statements, explanations and labels annotated by clinical domain experts from the Cancer Research UK Manchester and the Digital Experimental Cancer Medicine Team (Jullien et al., 2023). Each CTR has been pre-processed to include four sections: eligibility criteria, intervention, results, and adverse events. Each of these has been further subdivided into individual lines consistent with the structure of the CTR record. Each data sample includes:

- Statement: An annotated statement making a claim about about a CTR.
- Type: Whether the statement requires supporting information a single CTR or two CTRs to determine entailment.
- Section: The section of the CTR containing the information required to make a prediction

Additionally, a label is provided for each subtask. For subtask one this is a binary label of either *Entailment* or *Contradiction*. For subtask 2, this is a list of indices corresponding to relevant lines within a section which are necessary for determining entailment or contradiction of the statement. The data is split into train, validation, and test sets containing 1,700, 200, and 500 statements respectively.

3 System Overview

3.1 Pre-Processing

Using the JSON-formatted CTR and annotation files provided by the organizers, we create individual files for each of the sub-tasks consisting of newline-delimited JSON samples. For sub-task 1, each training and validation sample consists of a CTR premise, a statement, and a binary label (*Entailment* or *Contradiction*). For sub-task 2, each training and validation sample consists of a sequence of text from the relevant CTR section (split according to provided breaks) along with a binary label (*Relevant* or *Irrelevant*). For comparison type statements in which two CTRs are required to determine entailment (and thus, relevance), we prepend each text sequence in sub-task 2 samples with the string "first:" if it belongs to the first CTR and "second:" if it belongs to the second CTR, in an attempt to preserve an indication of CTR order.

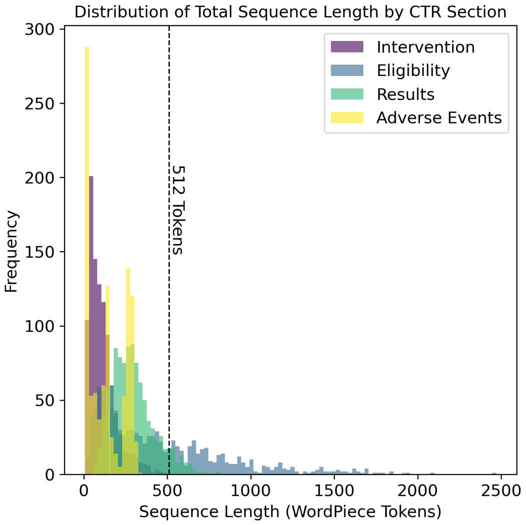

Figure 1: Histogram of sequence lengths for each CTR section. Note that these are lengths for individual CTR sections, which does not account for the length of statements (mean length: 16 tokens) and training samples with two CTR sections, both of which increase the effective input sequence length.

3.2 Model Architecture

For our foundation model we choose the GatorTron-BERT base model (Yang et al., 2022) using the pretrained weights released by the authors. This model is similar in architecture to BERT-large (Devlin et al., 2018) and consists of 24 layers, 16 attention heads per layer, and a hidden size of 1024, with a total parameter count of over 345 million. GatorTron was pretrained on a corpus of over 90 billion words which included de-identified clinical notes from the University of Florida Health System, PubMed articles, and Wikipedia articles, which was accomplished using a masked-language modeling objective as well as a sentence-order prediction objective, in which the model must predict the order of two segments of text (Yang et al., 2022). We additionally append a classification head to the base model for fine-tuning on both sub-tasks.

3.3 Domain Adaptation

To attempt to mitigate the effect of domain shift, we domain adapt our best performing fine-tuned models on a dataset in the target domain. The domain adaptation dataset is also compiled from CTRs from `https://clinicaltrials.gov/ct2/home`, with the constraints that the trials are completed, include results, and are cancer-related, resulting in approximately 9,000 CTRs. After filtering out trial IDs used in the Task 7 dataset, we parse the remaining XML CTR files to extract text from the four sections that appeared in the Task 7 dataset: Invervention, Eligibility Criteria, Results, and Adverse Events. The text from within each section is concatenated and then truncated to the maximum sequence length of our model, 512 tokens, to form the domain adaptation set. This dataset is then used to train the model on a masked-language modeling objective for a single epoch. The choice of domain adapting for one epoch is decided experimentally; we find that domain adaptation on this dataset beyond this amount adversely affect downstream performance.

3.4 Extraction and Classification Pipeline

Our approach to sub-task 1 can be viewed as a two-stage pipeline in which one model, the *extraction model*, extracts relevant text to justify a statement and another model, the *classification model* uses this text to determine the entailment relation. For the extraction model, we use the GatorTron model described in Model Architecture which has been fine-tuned on the sub-task 2 training set. For the classification model, we use a similar model that has been fine-tuned on the sub-task 1 training set with unfiltered CTR premises.

At test time, a CTR premise is split into lines as outlined in Pre-Processing, which are individually passed as input to the extraction model, along with the statement. Premise lines which are classified as relevant by the extraction model are concatenated, preserving source order, and passed on to the classification model, which predicts the entailment relation.

4 Experiments

4.1 Experimental Setup

We tokenize the datasets using WordPiece tokenization (Schuster and Nakajima, 2012) with the 50,000 token clinical vocabulary released alongside the GatorTron model[2] (Yang et al., 2022). Training and development splits provided by the organizers are used for both sub-tasks. All models are fine-tuned for 5 epochs, with a warmup ratio of 0.1, a learning rate of 5e-5, and a total batch size of 32. Cross-entropy is used as the loss function on both sub-tasks. All training was conducted on two NVIDIA A100 GPUs; fine-tuning took approximately 10 minutes for subtask 1 and 40 minutes for subtask 2, using full precision and data parallelism.

4.2 Submission and Evaluation

For our submissions, we choose the best performing model on the validation set out of the five epochs. System submissions are evaluated on F1-score, precision, and recall for both sub-tasks.

5 Results

5.1 Main Results

The results for each of the tested models can be seen in Table 1. For sub-task 1, we find that the extraction approach from 3.4 provides the largest advantage, with an absolute F1-score increase of 0.059 compared using full CTRs. This is likely due to the reduction of noise within the sample dataset as well as the reduction of sequence lengths to below the maximum model input length.

Results show a minor increase in performance for domain-adapted models for both sub-tasks with an absolute F1-score increase of 0.009 for sub-task 1 and 0.019 for sub-task 2 above the base models without domain-adaptation. This increase was not as significant as anticipated, which may indicate a relative similarity between the source domain and the target domain, though more analysis would be needed to verify this.

5.2 Annotation Artifacts

When providing the model only the statement without a CTR premise, it achieves an F1-score of

[2] Base model and vocabulary can be found at `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/gatortron_og`

	F1-Score	Precision	Recall
Task 1			
TF-IDF Baseline (Jullien et al., 2023)	0.502	0.486	0.520
Statement Only	0.584	0.602	0.568
BioBERT (Lee et al., 2020)	0.591	0.652	0.540
GatorTron-Base: Full Premise	0.637	0.520	0.820
GatorTron-Base Relevant-Only	0.696	0.648	0.752
GatorTron-Base: DA & Relevant-Only	0.705	0.654	0.764
Task 2			
BM25 Baseline (Jullien et al., 2023)	0.323	0.422	0.261
BioBERT (Lee et al., 2020)	0.794	0.795	0.792
GatorTron-Base	0.787	0.767	0.807
GatorTron-Base: DA	0.806	0.802	0.811

Table 1: Performance comparison of each approach on sub-tasks 1 and 2. Model with the highest F1-score for each sub-task is marked in black.

0.584, significantly better than the term frequency-inverse document frequency baseline F1-score of 0.502. Thus the model is able to infer an entailment relation, to some extent, from the domain-expert-annotated statements alone by leveraging *annotation artifacts* (Gururangan et al., 2018). These artifacts are a common occurence in natural language inference datasets, typically arising from imbalances in the sample distribution, like re-occuring words or patterns in entailment or contradiction statements (Gururangan et al., 2018). To determine whether entailment can be inferred simply from statement sequence lengths, we compare the sequence length distributions of both entailment and contradiction statements in Figure 2. While the contradiction distribution does skew more toward longer sequence lengths, it is unclear whether this difference is sufficiently large so as to be exploited for entailment recognition. More extensive studies are needed to identify the sources of these artifacts as well as their effect on model performance on this dataset.

5.3 Error Analysis

We individually checked the results on the validation set and find that our system tends to struggle with numerical reasoning, which constitutes a key element of this dataset. In particular, results on the validation set suggest that our system is fails to accurately perform comparison operations on non-discrete numbers.

6 Conclusion

In this paper we outlined our system for SemEval 2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data. We compare the effect of different training approaches on the performance of a large pretrained language model for the two sub-tasks. We find that while increased model size tends to lead to better performance, increased model size is not enough on its own to achieve strong performance on this dataset. We further demonstrate that continued pretraining on data in the target domain leads to a marked performance increase on the test set.

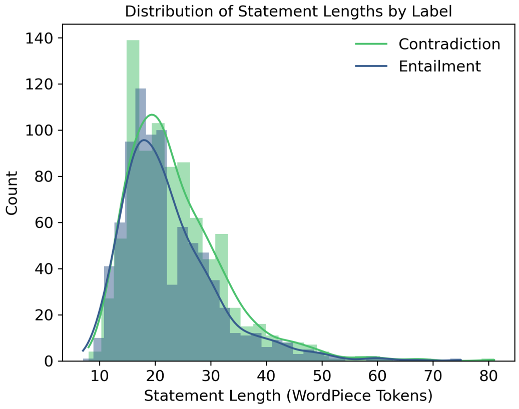

Figure 2: Histogram of sequence lengths for each statement grouped by label.

Due to the potential confounding effect of annotation artifacts, investigation into adversarial approaches for data augmentation on this dataset may be useful in the future (Nie et al., 2019). Furthermore, enhancing the numerical reasoning capabilities of language models (Ravichander et al., 2019; Geva et al., 2020) could benefit system performance on this dataset and warrants further research.

7 Acknowledgments

Clemson University is acknowledged for generous allotment of compute time on Palmetto cluster. We would also like to thank Dr. Varun Gangal for his help on this project as well as the reviewers for taking the time to review our manuscript and for providing valuable feedback.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. *arXiv preprint arXiv:2004.04487*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

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*.

Janice L Kwan, Lisha Lo, Jacob Ferguson, Hanna Goldberg, Juan Pablo Diaz-Martinez, George Tomlinson, Jeremy M Grimshaw, and Kaveh G Shojania. 2020. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ*, 370.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. 2021. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*.

Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.