I2R at SemEval-2023 Task 7: Explanations-driven Ensemble Approach for Natural Language Inference over Clinical Trial Data

Saravanan Rajamanickam and Rajaraman Kanagasabai

Agency for Science, Technology and Research (A*STAR)
Institute for Infocomm Research
Singapore 138632
{saravanan_rajamanickam, kanagasa}@i2r.a-star.edu.sg

Abstract

In this paper, we describe our system for SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data. Given a CTR premise, and a statement, this task involves 2 sub-tasks (i) identifying the inference relation between CTR - statement pairs (Task 1: Textual Entailment), and (ii) extracting a set of supporting facts, from the premise, to justify the label predicted in Task 1 (Task 2: Evidence Retrieval). We adopt an explanation driven NLI approach to tackle the tasks. Given a statement to verify, the idea is to first identify relevant evidence from the target CTR(s), perform evidence level inferences and then ensemble them to arrive at the final inference. We have experimented with various BERT based models and T5 models. Our final model uses T5 base that achieved better performance compared to BERT models. In summary, our system achieves F1 score of 70.1% for Task 1 and 80.2% for Task 2. We ranked 8th respectively under both the tasks. Moreover, ours was one of the 5 systems that ranked within the Top 10 under both tasks.

1 Introduction

Natural language inference (NLI) is the task of detecting inferential relationships between a premise text and a hypothesis text (MacCartney and Manning, 2009), which is considered fundamental in natural language understanding (NLU) research (Bowman et al., 2015). The objective is to determine whether hypothesis h is true ('entailment'), false ('contradiction'), or undetermined ('neutral') given the premise P. This task, formerly known as recognizing textual entailment (RTE) (Dagan et al., 2005) has long been a popular task among researchers. Moreover, contribution of datasets from past shared tasks (Dagan et al., 2009), and recent research (Bowman et al., 2015; Williams et al., 2018) have pushed the boundaries for this seemingly simple, but challenging problem.

NLI brings an opportunity to support the large-scale interpretation and retrieval of medical evidence. Understanding the contextual evidence will support personalized care for patients (Sutton et al.,2020), e.g., analysis of Clinical Trial Reports (CTRs). This is especially useful as, in the past few years, the number of publications of CTRs has increased exponentially and it has become impracticable for clinical practitioners to stay updated (DeYoung et al., 2020).

SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (Jullien et al., 2023) attempts to capture and investigate this opportunity via two shared tasks namely (i) Task 1: Textual Entailment: Given a statement, which make some type of claim about the information contained in one of the sections in the CTR premise, identify the inference relation (either entailment or contradiction) between CTR - statement pairs. (ii) Task 2: Given a CTR premise, and a statement, extract a set of supporting facts, from the premise, to justify the label predicted in Task 1.

Several NLI systems have been proposed in the literature. (Bowman et al., 2015, Romanov et al., 2015, Marelli et al., 2014, Khot et al., 2018, Williams et al., 2018, Ravichander et al.,2019, Nie et al., 2020) However, most of the successful

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1630–1635

systems typically assume short length premises and statements. In clinical trials data, while the statements are short, the premises can be much longer, even without considering comparison of CTRs. Thus, it is likely that the token size limit of even large language models is exceeded more often. This makes the tasks non-trivial. Moreover, clinical trial data contains complex, high volume of text and highly unstructured with distinct entities specific to clinical domains making the tasks challenging.

In our work, we adopt an explanations-driven NLI approach to tackle these challenges. Given a statement to verify, the idea is to first identify relevant evidence from the target CTR(s), perform evidence level inferences and then ensemble them to arrive at the final inference. The advantage is that the approach facilitates tackling both Task 1 and Task 2. We have experimented with several approaches using BERT (Devlin et.al. 2019) based models and T5 models (Raffel et al., 2020), and present the results obtained in Section 4. We observe that our ensemble approach using T5 base achieved the best performance in comparison.

In summary, our system achieves F1 score of 70.1% for Task 1 and 80.2% for Task 2. We ranked 8th respectively under both the tasks. Moreover, ours was one of the 5 systems that ranked within the Top 10 under both tasks. We conclude that our proposed approach is promising for NLI over Clinical text.

2 Background

2.1 Dataset:

The dataset is based on a collection of breast cancer CTRs (which are extracted from https://clinicaltrials.gov/ct2/home), statements, explanations, and labels annotated by domain expert annotators. Using various clinical domain experts, clinical trial organizers, and research oncologists from the Cancer Research UK Manchester Institute and the Digital Experimental Cancer Medicine Team, statements and evidence are generated for this task.

In total, there are 2,400 statements split evenly across the different sections and classes. Each Clinical Trial Report (CTR) consists of 4 sections (Eligibility criteria, Intervention, Results, and Adverse events). Each CTR may contain 1-2 patient groups, called cohorts or arms. These groups may receive different treatments or have different baseline characteristics. It consists of two sub tasks:

- Task 1: Textual Entailment
- Task 2: Evidence retrieval

2.2 Task 1: Textual Entailment

Each instance for task 1 contains 1-2 CTRs, a statement, a section marker, and an entailment/contradiction label. Task 1 is to determine the inference relation (entailment vs contradiction) between CTR - statement pairs.

The annotated statements are sentences with an average length of 19.5 tokens. The statements may make claims about a single CTR or compare two CTRs. Figure 1 illustrate the Task 1 with an input containing Statement, Label and Section extracted from the dataset.

Input:
Statement:
The primary trial and the secondary trial both used MRI for their interventions.
Label: Entailment
Section: Intervention

Output: On seeing both primary and secondary trail reports under intervention section, model needs to infer the relationship as “entailment”.

Figure 1. Task 1 Illustration
(see NCT02429427 for the original report)

2.3 Task 2: Evidence retrieval

Given a CTR premise, and a statement, output a set of supporting facts, extracted from the premise, necessary to justify the label predicted in Task 1. Figure 2 shows an example for Task 2 containing Statement, Label and Section name and details extracted from primary trial and secondary trial.

Input:
Statement:
More than 1/3 of patients in cohort 1 of the primary trial experienced an adverse event.
Label Contradiction
Section Adverse events
Primary Trial
Adverse Events 1:
- Total: 69/258 (26.74%)
- Anaemia 3/258 (1.16%)
- Febrile neutropenia 13/258 (5.04%)

- Neutropenia 5/258 (1.94%)
- Thrombocytopenia 1/258 (0.39%)
- Atrial fibrillation 0/258 (0.00%)
- Mitral valve incompetence 1/258 (0.39%)
- Pericardial effusion 0/258 (0.00%)
- Sinus tachycardia 0/258 (0.00%)
- Abdominal pain 3/258 (1.16%)
- Abdominal pain upper 1/258 (0.39%)
- Colitis 1/258 (0.39%)

Output:

As per the primary trial results, the total number of patients who have experienced an adverse event is less than 1/3 of patients in cohort 1. Task 2 will provide a set of relevant supporting facts to justify the label predicted in Task 1.

Figure 2. Task 2 Illustration

3 System Overview

Several NLI systems have been proposed in the literature review using large general domain dataset such as SNLI, MNLI (Bowman et al., 2015, Williams et al., 2018) and medical domain dataset MEDNLI(Romanov et al., 2015). However, most of the successful systems typically assume short length premises and statements. In clinical trials data, while the statements are short, the premise can be quite long, even without considering comparison of CTRs. Thus, it is likely that the token size limit of even large language models is exceeded more often. This makes the tasks non-trivial. Moreover, the data requires heavy use of domain knowledge making the tasks challenging.

Figure 3. System Workflow

Broadly, we adopt an explanation driven NLI approach whereby, given a statement (or hypothesis) to verify, the idea is to first identify relevant evidence from the target CTR(s), perform evidence level inferences and then ensemble them to arrive at the final inference (Figure 3). Reducing the inference to evidence level texts facilitates the application of standard transformer models. We explain the steps in detail below.

3.1 Evidence Classification:

Given a CTR, first we identify evidence (split lines as indexed in a CTR section) relevant for the NLI task. The primary and secondary evidence texts tagged in a CTR (from Task 2) are extracted and labeled as *'relevant'* and others in the respective sections are labeled *'irrelevant'*. We formulate a binary sentence pair classification problem with the statement as *sentence1* and an evidence text as *sentence2*. We apply T5 (Raffel et al., 2020) as the encoder-decoder model to generate the classifications given the sentence pairs.

The evidence is extracted from the dataset, and a training set is prepared using the processed input as *'mnli hypothesis: <statement> premise: <evidence text>'* and the output is 'relevant' or 'irrelevant', as explained above. We measure precision and recall using the exact matching ratio.

3.2 Evidence-level NLI:

Evidence classified in the previous step are used to perform a 3-way NLI as below. For evidence classified as 'relevant', we assign target labels from Task 1 ('Entailment' or 'Contradiction'), and label irrelevant sentences as 'Neutral'.

We train a second T5 model as below. A training set is prepared using the processed input as *'mnli hypothesis: <statement> premise: <evidence text>'* and the output is *'Entailment'*, *'Contradiction'* or *'Neutral'*, as explained above. We measure precision and recall using the exact matching ratio.

Now, for sub task, (i) Task 1: Textual Entailment, we ensemble the evidence-level NLI labels and predict final inference based on the ratio of *'Contradiction'* labels over *'Entailment'* labels. The prediction is *'Entailment'* if the ratio is less than a threshold and there is at least one evidence labeled *'Entailment'*, otherwise the prediction is 'Contradiction'. In our experiments, max pooling is to determine the relationship between CTR and statement pairs.

For Task 2: Evidence Retrieval, we predict the evidence indices corresponding to all evidence classified as 'relevant'. Utilizing the multi-task learning objective of T5 model, we have used the same model to fine tune the model to retrieve the set of supporting facts from the CTR premise.

Our system is evaluated using standard evaluation metrics - precision, recall and F1-score.

4 Experiments

In the NLI4CT challenge dataset, there are 2400 instances provided, and the labels are evenly split across train/dev/test (1700/200/500) examples.

Our model, as described in Section 3, was initialized from the pre-trained T5 base trained on a variety of general text on MLM scheme. We have used this as a starting point to fine tune the model for the given dataset and adjusted the hyper-parameters based on the best performance. We have also used the feature relevancy to categorize the relevant sentences. During T5 training, we set the number of beams as 50 and the number of returned sequences as 5. We randomly split the instances into 80% training and 20%, repeated validation five times and report the average performance.

The experiments were executed on NVIDIA-GeForce RTX 2080 series with eight cores of GPU machines with 8*12 GB of memory for all our experiments. Also, to train T5 large models, we have used NVIDIA-GeForce Tesla V100 series SXM2-32GB with 5 cores of GPU machines. Models were trained for 3-5 hours for training and reasoning. The pretrained weights for the transformers prior to fine-tuning were from the HuggingFace NLP Library.

The results are presented below, where performance metrics are averaged over 10 runs and quoted in % for easier interpretation, unless stated otherwise.

4.1 Approach I with Baseline models (Table 1)

In our first approach for Task 1, we considered the premise and statement as two text chunks and attempted a naïve NLI approach by straight-forward fine-tuning on the NLI4CT training data. We experimented with the baseline language models BERT (Devlin et.al. 2019), RoBERTa(Liu et.al., 2019) and DeBERTa(He et.al., 2021). Additionally, we also considered sparse attention models such as Bigbird (Zaheer et al., 2020) and Longformer (Beltagy et.al., 2020) models to capture long range dependencies for longer documents. However, we did not achieve comparable results with these full attention transformers-based models.

For Task 1 baseline experiments, we have re-implemented the fine-tuned BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) base version and used *[CLS] CTR premise [SEP] statement [SEP]* as input to the transformers to predict the logical relation. Also, repeated the experiment with only relevant sentences specified as supporting facts for Task2.We have used *[CLS] premise - (r1,r2..rn) [SEP] statement [SEP]* where {r1,r2..rn} are relevant supporting facts. Models trained with relevant sentences performed better than the model trained with whole CTR premise. Our models are trained end-to-end using AdamW optimizer with the decay rate of 0.9 and learning rate of 5e-6 for our BERT and RoBERTa base version.

Models	Dev Accuracy (%)
BERT-base	58.5%
RoBERTa-base	50.5%
DEBERTA-base	51.5%
t5-base	62.2%

Table 1: Comparison of Accuracy against Baseline models for Task 1

The performance of T5 models in Task 1 is comparatively better than BERT based models, and so we decided to adopt T5 in subsequent studies.

As a key observation, in error analysis, we noted that the model was unable to infer effectively mainly in view of longer texts. Hence, we decide on an alternate approach exploiting the evidence texts.

4.2 Approach II with T5 models (Table 2)

In this approach, we treat the evidence texts as possible explanations of the inference and use them to drive the NLI process. In particular, given a statement to verify, we first identify relevant evidence from the target CTR(s), perform evidence level inferences and then ensemble them to arrive at the final inference.

To provide context, we prefix each evidence text with the Section title, and whether it belongs to primary or secondary trial, as illustrated in Figure 4.

evidence_	evidence_text
24	Primary trial Eligibility - Active alcohol or drug abuse
25	Primary trial Eligibility - Other malignancy within the past 5 years
0	Secondary trial Eligibility - Premenopausal women 55 years of age or younger with regular menstrual cycles (at least four cycles in the last six months). Women with fewer than 4 menses in the last 6 months or who have had a hysterectomy with ovaries intact will be considered premenopausal if FSH level < 20.
1	Secondary trial Eligibility - Women with breast density 25% (scattered fibroglandular densities or greater) are eligible.
2	Secondary trial Eligibility - Prior Treatment
3	Secondary trial Eligibility - Patients who are currently receiving hormone replacement therapy (estrogen or progesterone); or are taking tamoxifen or raloxifene are not eligible. Women who have taken these medications must have stopped for at least 4 months prior to study entry.
4	Secondary trial Eligibility - Topical estrogen (eg, transdermal patches and vaginal estrogens) is allowed.

Figure 4. Illustration of Prefixed Evidence Texts

We also experimented by adding 'Patient Group' subsection titles, but this did not improve the final performance.

Next, by using the Evidence classification and Evidence-level inference steps outlined in Section 3, we trained two T5 models, one each for Task1 and Task2 and obtained results as in Table 2.

Task 1			
	F1 Score	**Precision**	**Recall**
t5-base	62.91%	59.29%	67.00%
t5 large	68.35%	59.12%	81.00%
Task 2			
t5 base	85.08%	82.78%	87.51%
t5 large	82.87%	79.92%	86.04%

Table 2 Comparison of t5 base and t5 large model for Task1 and Task2

We observe that the new approach outperformed all baseline models.

4.3 Approach III with Post-tuned T5 model

Though Approach II had better performance, our error analysis revealed that it was limited in situations where no relevant evidence is found. This could happen either when the evidence section contains substantial text, or the CRT uses too much domain specific terminologies. To handle these cases, we extracted CTR's that had empty primary or secondary evidence in the validation set and reviewed manually. We derived a set of rules to handle exceptions and negations. We performed a Task1 run using Approach II and reviewed the empty results cases. 'Entailment' was assigned a default inference. It was toggled based on careful human judgment by referring to the exception rules. This resulted in improving our final Task1 performance above F1 score of 70%.

5 Results

Table 3 shows the final model performance and best results obtained for Task 1 & 2.

Sub task	**F1**	**Precision**	**Recall**
Task1	70.1% (8)	55.0% (17)	96.8% (4)
Task2	80.2% (8)	79.7% (6)	80.7% (10)

Table 3 Model Performance for Task1 and Task2

For Task 1, our system achieved low precision and high recall compared to other teams that participated in the task. The model performance in Task 2 seems to be better compared to all metrics.T5 models works better for both Task 1 and Task 2.

6 Conclusions

In this paper, we described our system for SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data. We adopted an explanation driven NLI approach to develop a methodology to tackle both the tasks. We have experimented with various BERT based models and T5 models, and our final model uses T5 base that achieved better performance. Our system achieves F1 score of 70.1% for Task 1 and 80.2% for Task 2. We ranked 8th respectively under both the tasks. Moreover, ours was one of the 5 systems that ranked within the Top 10 under both tasks.

We noted that the *T5 base model* for natural language inference task performs reasonably well for clinical trial dataset. Some interesting research questions for further investigation are: 1) Generative text to text framework T5 models could perform well for NLI along with arithmetic reasoning. 2) Pre-training data with domain specific clinical corpora might increase performance. Using pre-trained models trained on large biomedical corpora such as SciFive (Phan et.al., 2021) and BioBERT (Lee et.al., 2019) models are steps along this direction.

References

Iz Beltagy, Matthew E. Peters, Arman Cohan. 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of EMNLP.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Machine Learning Challenges Workshop, pages 177–190. Springer.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering, 15(4):i–xvii.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of

deep bidirectional transformers for language understanding. In NAACL-HLT.

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/2005.0417

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: decoding-enhanced bert with disentangled attention. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

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.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. CoRR, abs/1901.08746.

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. CoRR.

Bill MacCartney and Christopher D Manning, 2009. An extended model of natural logic. In Proceedings of the eighth international conference on computational semantics, pp. 140–156. Association for Computational Linguistics.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. arXiv preprint arXiv:2106.03598.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4885–4901, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *In Journal of Machine Learning Research.*

Abhilasha Ravichander, Aakanksha Naik, Carolyn Penstein Rosé, and Eduard H. Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. CoRR, abs/1901.03735.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Martin Schmitt and Hinrich Schütze. 2019. SherLIiC: A Typed Event-Focused Lexical Inference Benchmark for Evaluating Natural Language Inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 902–914, Florence, Italy. Association for Computational Linguistics.

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):1–10

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of NAACL.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, C. Alberti, S. Ontan˜on, ´ Philip Pham, Anirudh Ravula, Qifan Wang, L. Yang, and A. Ahmed. 2020. Big bird: Transformers for longer sequences. ArXiv, abs/2007.14062.