JUST-KM at SemEval-2023 Task 7: Multi-evidence Natural Language Inference using Role-based Double Roberta-Large

Kefah Alissa and **Malak Abdulah**
Jordan University of Science and Technology
Irbid, Jordan
{kfalissa20, mabdullah}@just.edu.jo

Abstract

In recent years, there has been a vast increase in the available clinical data. Variant Deep learning techniques are used to enhance the retrieval and interpretation of these data. This task deployed Natural language inference (NLI) in Clinical Trial Reports (CTRs) to provide individualized care that is supported by evidence. A collection of breast cancer clinical trial records, statements, annotations, and labels from experienced domain experts. NLI presents a chance to advance the widespread understanding and retrieval of medical evidence, leading to significant improvements in connecting the most recent evidence to personalized care. The primary objective is to identify the inference relationship (entailment or contradiction) between pairs of clinical trial records and statements. In this research, we used different transformer-based models, and the proposed model, "Role-based Double Roberta-Large," achieved the best result on the testing dataset with F1-score equal to 67.0%

1 Introduction

Breast cancer clinical trial reports (CTRs) summarize the results of clinical trials conducted to test new treatments or interventions for breast cancer. These reports typically include information on the study design, patient population, treatment regimens, outcomes (such as response rates and side effects), and overall conclusions. These reports provide a comprehensive and transparent account of the trial results to the scientific community and the general public to inform future research and clinical practice. The results of breast cancer clinical trials play a critical role in the development of new treatments and in improving outcomes for patients with breast cancer(McDonald et al., 2016). The number of published (CTRs) has grown significantly in recent years, with over 10,000 CTRs for breast cancer alone. As a result, it has become increasingly more work for clinical practitioners to keep up with all the current literature and provide patient care based on the latest evidence (DeYoung et al., 2020).

Natural Language Inference (NLI) has the potential to enhance the analysis and interpretation of CTRs by providing a more automated and consistent approach to extracting information and drawing inferences from these reports(Agrawal et al., 2019). Multi-evidence Natural Language Inference (NLI) is a task in natural language processing that involves determining the relationship between two pieces of text, specifically whether the second text is an entailment, contradiction, or neutral concerning the first text. In multi-evidence NLI, multiple pieces of evidence are used to make the determination instead of relying on a single text. This can be useful when the relationship between the two texts needs to be clarified or additional context is needed to make an accurate determination(Storks et al., 2019). NLI has been approached using various methods, including symbolic logic, knowledge bases, and neural networks. Recently, it's become a crucial testing ground for methods using distributed word and phrase representations. These distributed representations excel in capturing similarity-based relations and have successfully modeled basic dimensions of meaning such as evaluative sentiment(Socher et al., 2013). Following the substantial success of deep learning (DL) techniques in various artificial intelligence tasks, researchers in the field of natural language processing (NLP) have begun to develop DL-based models to analyze patterns in the natural language data produced by humans(Otter et al., 2020).

This research aims to build a deep learning model to identify the inference relationship (Entailment or contradiction) between pairs of CTR statements (textual relationship - prompt). The rest of this paper goes as follows: Section II reviews the recent research works related to this study. Section III describes our methodology. Section IV reports

1
Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 447–452

the results by discussing the proposed model and comparative performance. Finally, the conclusion is in Section V.

2 Related Work

Deep learning has emerged as a powerful tool for analyzing and interpreting complex medical data and has been widely studied in recent years (Alissa et al., 2022; Abedalla et al., 2021; Alharahsheh and Abdullah, 2021). This section presents a brief review of the related work in this field. The authors(Wang and Jiang, 2015) presented a new LSTM architecture for natural language inference (NLI) that builds on top of a previous neural attention model. The proposed model performs word-by-word matching of the hypothesis and the premise using a match-LSTM. The LSTM emphasizes important word-level matches and remembers critical mismatches to predict the relationship label. The model is evaluated on the Stanford Natural Language Inference (SNLI) corpus and achieves an accuracy of 86.1%, surpassing state-of-the-art.

The authors(Chen et al., 2016) presented a new state-of-the-art approach for the Stanford Natural Language Inference Dataset with a high accuracy of 88.6%. Instead of using complex network architectures, they demonstrate that a carefully designed sequential inference model using chain LSTMs outperforms previous models. They achieve further improvement by incorporating recursive architectures in local inference modeling and inference composition. Additionally, syntactic parsing information contributes to the best results, even when added to a powerful model. The authors(Zhang et al., 2020) proposed a new language model, SemBERT, that combines BERT with pre-trained semantic role labeling to explicitly capture contextual semantics. SemBERT is easy to fine-tune for specific tasks while retaining BERT's usability. It outperforms BERT and sets new state-of-the-art results on ten reading comprehension and language inference tasks. The authors(Ghaeini et al., 2018) presented a novel deep learning architecture, DR-BiLSTM, for the Natural Language Inference task. They are using the Stanford NLI dataset. After an enhanced preprocessing step, they achieved new state-of-the-art scores by the DR-BiLSTM model.

DR-BiLSTM used a different approach from other existing methods, which used a simple reading technique. DR-BiLSTM can model the relationship between a hypothesis and a premise while encoding and inference in an efficient way using dependent reading. Also, the ensemble technique is used to merge the models to enhance predictions. The authors(Wu and Huang, 2022) proposed a multi-branch network that combines knowledge and context information to improve performance. The network has two branches, one for context information that uses multi-level dynamic assisted attention to construct interaction between sentence pairs and another for knowledge information that employs a Knowledge-based Graph Attention Network (K-GAT) to capture structural information of knowledge and uses an attention mechanism for interaction. Additionally, a relation branch captures context and knowledge relations between sentence pairs. The model uses five semantic dependencies-based knowledge types to minimize redundant external knowledge. The experiments show solid competitive results on three popular NLI datasets.

3 Methodology

This section outlines our methodology and proceeds as follows: Firstly, the dataset for the task is outlined. Then, the different preparing dataset approaches from the CTRs are described. Finally, we described the proposed model approach to identify the inference relationship (entailment or contradiction) between pairs of clinical trial records and statements.

3.1 Dataset and Task description

This task involves compiling breast cancer Clinical Trial Reports (CTRs) obtained from[1]. The CTRs have been annotated by domain experts and summarized into four sections: Eligibility Criteria, Intervention, Results, and Adverse Events. The eligibility criteria is a set of criteria for patients to participate in the clinical trial that is established, but the Intervention contains the details regarding the type, amount, frequency, and length of treatments under provided examination, the results section contains the number of participants, performance indicators, units of measurement, and the recorded outcomes, the last section Adverse Events contains the symptoms and the noticed indications throughout the clinical trial. The annotated statements in this task are sentences, averaging 19.5 tokens in length, that express claims about information found in one of the CTR sections. These claims can pertain to a single CTR or compare two CTRs.

[1]https://clinicaltrials.gov/ct2/home

The SemEval-task 7 (Jullien et al., 2023) competition has provided three JSON files (train, dev, and test data). The files offer the Premise-Statement-Evidence details described in Tabel 1, and the CT JSON folder comprises the complete set of CTRs stored as individual JSON files.

Table 2 displays a sample of two sentences from the dataset with the extracted evidence from the provided CTR depending on Its section and type. Unfortunately, the two sentences seem too close, and although they share the same primary evidence, they have different inferences, "Entailment" and "Contradiction". This can lead us to the complexity of the dataset.

3.2 Dataset statistics

Table 3 includes statistics about the number of samples in each set (train/validation). We can conclude can that data is balanced through this table. And in the testing dataset consists of 500 CTRs samples.

3.3 Dataset preparing

First, we extracted the evidence from the CTR depending on the CTR name, section, and evidence indices. Then, we concatenated the evidence to form one text. Then, we removed the NaN values. After that, we concatenated the secondary and primary evidence for the samples of type comparison. Finally, we extracted the sentences that related to the same conditions, including the (type, section, and evidence indices from both primary and secondary) but have different Labels as the example in Table 2.

At the end of this step, we prepared two datasets. The first one is (EvidenceSen) which contains all the sentences in the dataset with Its extracted evidence Labeled by "Contradiction" or "Entailment". And the (ContradictionSens) which contains the sentences grouped by CTR file name (primary when the type is single) and (primary and secondary when the sample type is "comparison"), each sample contains a sentence with an opposite sentence that is extracted from a sample has the same evidence but labeled with a different Labels. While preserving the Ids on each sample.

3.4 Role-based Double Roberta-Large

Transformer-based models have achieved state-of-the-art results in various natural language processing tasks, including text summarization, sentiment analysis, question answering, natural language inference, and others (Antoun et al., 2020). In this study, we utilized RoBERTa. It (short for Robustly Optimized BERT Pretraining Approach) is a pre-trained language model developed by Facebook AI Research (FAIR) in 2019. It is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture. It has been trained on a diverse range of internet text to achieve state-of-the-art results on various NLP tasks. RoBERTa outperforms BERT and has become a popular choice for fine-tuning specific NLP tasks(Liu et al., 2019)

Our proposed model, "Role-based Double Roberta-Large," is shown in Figure.1 consists of two RoBERTa-Large models. The first model was trained on the prepared data set, which included the sentence and evidence classified by "Contradiction" or "Entailment" which is called the (EvidenceSen) dataset. The second model was trained on the (EvidenceSen) and the (ContradictionSens) data set. First, we checked if the sample contains a relevant sample in the (ContradictionSens) dataset. If not, we relied on the first model prediction with a threshold of 0.40% (if the softmax probability is greater than 40%, we predict it as "Entailment", otherwise, we predicted it as a "contradiction"). But if the sample has a sample linked in (ContradictionSens). We check if Model 2 predicts it as a "Contradiction", then we set the sample with the highest probability as "Entailment" and the other related sample as a "Contradiction," regardless of the probability of the first model.

3.5 Evaluation metrics

Evaluation metrics in deep learning are used to measure the performance of a model by assessing its accuracy and robustness in predicting outcomes. The most common evaluation metrics include accuracy, precision, recall, and F1 score. Each metric has strengths and weaknesses, so it is important to choose an appropriate evaluation metric for a given problem. In this task, we used the F1 score as the evaluation metric to measure the model's performance.

- **Accuracy :** the is the most common metric used to measure performance in machine learning models, and it simply measures the number of correct predictions made by a model as a percentage of all predictions made.
- **Precision:** This metric looks at how accurate a model's positive predictions are, and it measures the proportion of true positives divided by all positive predictions made by the model.

Table 1: The description of Premise-Statement-Evidence information

Key Name	Description
UUID	The initial component of the Premise-Statement pair.
Type	The entry can have two options: "Single" or "Comparison". In the case of "Single", there is only one trial, the Primary trial, and the statement will only relate to this trial, so all the proof will be found in the Primary CTR. When it is "Comparison", there are two trials, the Primary and Secondary trial (as referred to in the statements), and the statements refer to both trials, hence the evidence must be obtained from both CTRs
Section_id	The entry can have one of four possible values: "Eligibility criteria", "Intervention", "Results", or "Adverse events".
Primary_id	The "Primary_id" entry holds the identifier of the primary CTR.
Secondary_id	The "Section_id" entry can have one of four possible values: "Eligibility criteria", "Intervention", "Results", or "Adverse events".
Statement	The entry comprises a string of the annotated statement
Label	The "Label" entry can have one of two values: "Entailment" or "Contradiction".
Primary_evidence_index	The "Primary_evidence_index" entry holds a list of indices for the tagged entries in the specific section of the Primary CTR that serve as evidence.
Secondary_evidence_index	Entry holds a list of indexes referencing entries in the corresponding section of the Secondary CTR that have been designated as evidence if the "Type" entry is "Comparison."

Table 2: A sample of dataset

Sentences	Primary Evidence Index	Type	Section	Label
Adult Patients with histologic confirmation of invasive bilateral breast carcinoma (T3 N1 M0) are eligible for the primary trial.	[0]Inclusion Criteria: [1]Patients with histologic confirmation of invasive breast carcinoma. [3]Patients greater than or equal to 18 years. [4]Patients should have T1N1-3M0 or T2-4 N0-3M0. [5]Patients with bilateral breast cancer are eligible.	Single	Eligibility	Entailment
Adult Patients with histologic confirmation of invasive bilateralbreast carcinoma (T1 N1 M1) are eligible for the primary trial.		Single	Eligibility	Contradiction

Table 3: The CTRs Dataset statistics

Class	Training	Validation	Total
Contradiction	850	100	950
Entailment	850	100	950
Total	1700	200	1900

- **Recall:** Also known as sensitivity or true positive rate, this metric looks at what fraction of actual positives were correctly identified by the model out of all possible positives in the data set.
- **F1 Score:** This overall measure combines precisions and recalls.

4 Results and Discussion

We conducted experiments with various models to identify the most suitable ones for our task. The models we tried include BERT, RoBERTa-base, RoBERTa-large, TF-IDF, and our proposed model, "Role-based Double Roberta-Large". The results of our experiments are summarized in Table 4. We noticed that the proposed model achieved the best score, which equals 67.0%. Therefore, we fine-tune the hyperparameter for our proposed model. To ensure that the model is achieved the best results. Different hyperparameters have been tuned. And the best results are achieved by the below Table 5 hyper-parameters for both RoBERTa-large models in our proposed model, "Role-based Double Roberta-Large".

Table 4: Results of different transformers model in the testing phase

Model	F1 Score
TF-IDF	0.5702
BERT	0.6347
RoBERTa-base	0.6564
RoBERTa-large	0.6612
Role-based Double Roberta-Large	0.67

We observed that the performance of a Natural Language Inference (NLI) task using a large lan-

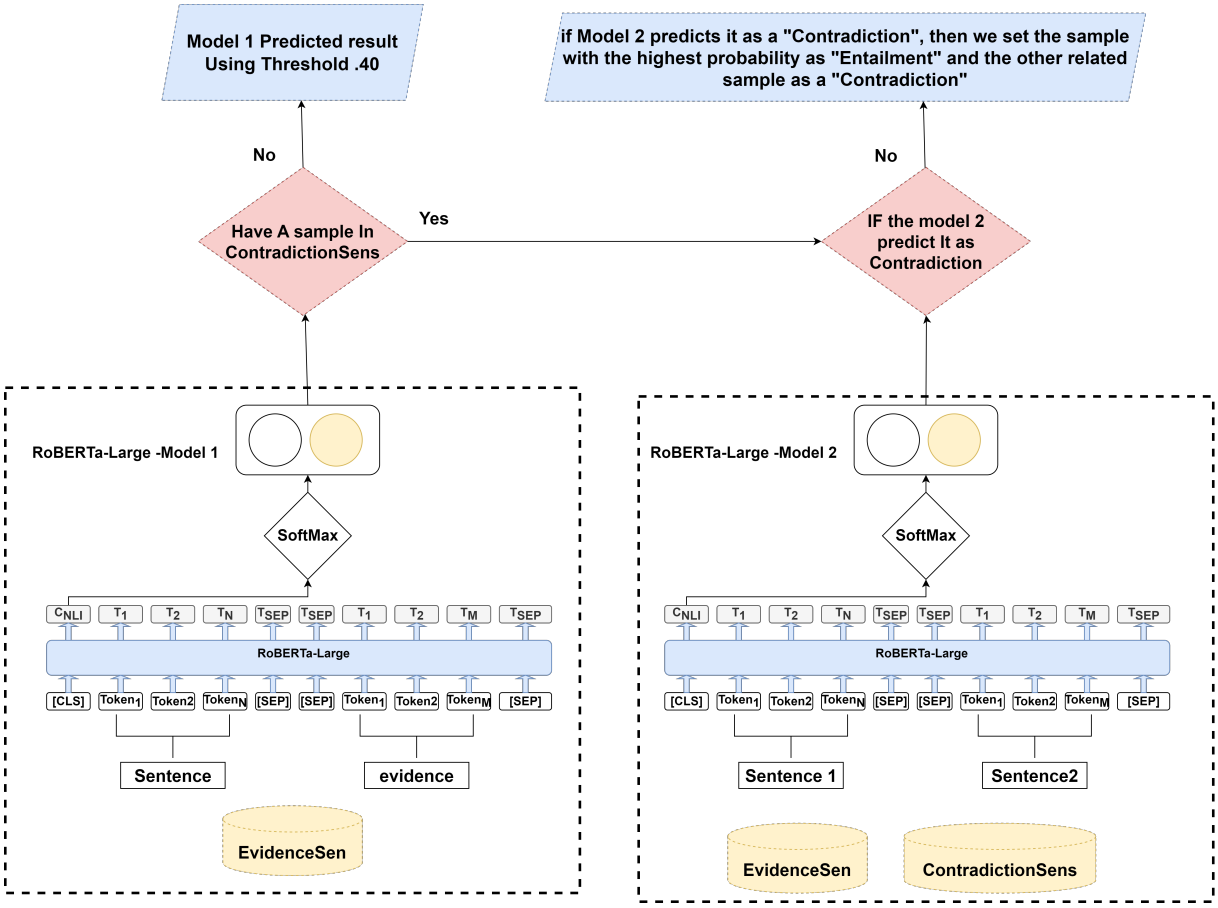

Figure 1: Role-based Double Roberta-Large

guage model like RoBERTa could be affected by the training data size. Smaller training datasets can lead to overfitting, where the model learns the training examples too well and needs to generalize better to new examples. This can result in poor performance on the NLI task. Additionally, RoBERTa is trained on a large corpus of diverse text. It is fine-tuned on specific tasks, so having a small dataset might need to provide more examples to fine-tune the model effectively. And the nature of CTRs data that is sensitive and need expert annotation makes the use of general augmentation techniques not recommended.

Table 5: Best Hyper parameter from the Proposed Model

Hyper parameter	Roberta-Large Model 1	Roberta-Large Model 2
Learning rate	4e-5	6e-6
Max sequence	64	128
Batch size	16	64
Epochs	5	3
early stopping metric	mcc	mcc

5 Conclusion

In conclusion, this research has explored the task of Natural Language Inference and its applications in NLP. Through the analysis of current state-of-the-art models and evaluation metrics, it has been shown that NLI is a challenging and complex task that requires modeling both lexical and semantic information.

This term paper has examined using the pre-trained transformer models for Natural Language Inference. Pre-trained transformer models have shown better results than other techniques in NLI tasks. Our experiments with the RoBERTa model on various NLI datasets demonstrate its effectiveness in classifying CTR sentence pairs into entailment or contradiction relationships. The results highlight the potential of RoBERTa for solving NLI problems. The Natural Language Inference (NLI) task is considered difficult due to the inherent ambiguity of language, the complexity of reason-

ing required to determine the relationship between two sentences, and the variability of sentence structures and relationships. Despite the challenges and limitations faced by NLI, it is clear that the task is of great importance for NLP and has numerous potential applications in areas such as text classification, question answering, and dialogue systems. Therefore, it is crucial for the research community to continue to advance the development of NLI models and explore new approaches for improving their performance.

References

Ayat Abedalla, Malak Abdullah, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. 2021. Chest x-ray pneumothorax segmentation using u-net with efficientnet and resnet architectures. *PeerJ Computer Science*, 7:e607.

Anumeha Agrawal, Rosa Anil George, Selvan Suntiha Ravi, Sowmya Kamath, and Anand Kumar. 2019. Ars_nitk at mediqa 2019: Analysing various methods for natural language inference, recognising question entailment and medical question answering system. In *Proceedings of the 18th BioNLP workshop and shared task*, pages 533–540.

Yara E Alharahsheh and Malak A Abdullah. 2021. Predicting individuals mental health status in kenya using machine learning methods. In *2021 12th International Conference on Information and Communication Systems (ICICS)*, pages 94–98. IEEE.

Kefah Alissa, Rasha Obeidat, Samer Alqudah, Rami Obeidat, and Qusai Ismail. 2022. Performance evaluation of cnn-based transfer learning for covid-19 pneumonia identification with various levels of layer partial freezing. In *2022 International Conference on Engineering & MIS (ICEMIS)*, pages 1–8. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

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*.

Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. *arXiv preprint arXiv:1802.05577*.

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*.

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*.

Elizabeth S McDonald, Amy S Clark, Julia Tchou, Paul Zhang, and Gary M Freedman. 2016. Clinical diagnosis and management of breast cancer. *Journal of Nuclear Medicine*, 57(Supplement 1):9S–16S.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.

Huiyan Wu and Jun Huang. 2022. Network based on the synergy of knowledge and context for natural language inference. *Neurocomputing*, 512:408–419.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.