Investigation of the Role of Context in Evidence Retrieval for Natural Language Inference from Clinical Trials

Adrita Anika †, **Md Messal Monem Miah, Tracy Hammond**

Department of Computer Science
Texas AM University
College Station, Texas, USA
adrita.anika@tamu.edu

Abstract

This paper explores the significance of context in evidence retrieval for natural language inference (NLI) from clinical trials. The study evaluates the performance of the state-of-the-art (SOTA) base models in retrieving evidence, considering both contextual and non-contextual scenarios. Various approaches for incorporating context are employed and their impact on model performance is systematically analyzed. Specifically, this study investigates the models' precision, recall , F1 scores and determine the optimal approach for integrating context. By elucidating the role of context through rigorous evaluation, this study provides valuable insights for NLI from clinical trials.

Introduction

Clinical trial Reports (CTRs) represent a category of medical research inquiries that are centered on investigations involving human subjects. CTRs are essential investigations in medical research that assess the effectiveness of new therapies compared to established treatments[1]. The dissemination of clinical trial findings often takes the form of lengthy natural language articles. Additionally, the volume of CTRs being published has experienced a significant surge over the past few years. The increasing number of CTRs and their lengthy publications pose challenges for healthcare professionals, who struggle to keep up with the vast amount of information. This situation hampers the delivery of personalized care based on reliable evidence (DeYoung et al. 2020). NLI can revolutionize the integration of medical evidence on a large scale, transforming the way we connect the latest findings to support personalized care (Sutton et al. 2020). NLI involves determining the logical relationship between a hypothesis, representing a physician's statement, and a premise, representing the CTR text. This logical relationship can be entailment, contradiction or neutral. To mention, a premise is a statement that serves as the foundational basis for an argument or the underpinning for a conclusion. Conversely, a hypothesis is a proposition or an assumption introduced to provide support to an argument. A statement is classified as a "contradiction" if the hypothesis is not implied by the premises.

Hypotheis	In total cohort 1 of NCT01439282 recorded 20% more adverse events than cohort 2.
Premise	Adverse Events 1: Total: 14/67 (20.90%), Febrile neutropenia 21/67 (1.49%), Leukopenia 21/67 (1.49%), Neutropenia 21/67 (1.49%), Macular hole 21/67 (1.49%), Diarrhoea 22/67 (2.99%), Abdominal pain 21/67 (1.49%), Abdominal pain upper 21/67 (1.49%)", Enteritis 21/67 (1.49%), Gastritis 21/67 (1.49%), Nausea 21/67 (1.49%), Vomiting 21/67 (1.49%), Pneumonia 21/67 (1.49%), Adverse Events 2: Total: 1/10 (10.00%), Febrile neutropenia 20/10 (0.00%) Leukopenia 20/10 (0.00%), Neutropenia 20/10 (0.00%), Macular hole 20/10 (0.00%), Diarrhoea 20/10 (0.00%), Abdominal pain 20/10 (0.00%), Abdominal pain upper 20/10 (0.00%), Enteritis 20/10 (0.00%), Gastritis 20/10 (0.00%), Nausea 20/10 (0.00%), Vomiting 20/10 (0.00%), Pneumonia 20/10 (0.00%)

Table 1: Example illustrating the retrieval of evidence, demonstrating the potential benefits of considering a larger context

Conversely, a statement is categorized as an "entailment" if it is implied by the premises, and considered neutral in other cases. By effectively retrieving evidence from CTRs, we can identify the crucial lines of text responsible for the decision of contradiction or entailment, filter out irrelevant information,

[1]https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies

and enhance model interpretability (Mishra and Sachdeva 2020). This approach of retrieving evidences simplifies the process by providing physicians with the relevant evidence sentences instead of overwhelming them with the entire CTR.

Recent studies have addressed the challenges of evidence retrieval from clinical trials. The authors (Mahendra, Spina, and Verspoor 2023a) approached this task as a document retrieval and sentence similarity problem, using BM25 (Robertson, Zaragoza et al. 2009) and Word Mover Distance algorithms (Kusner et al. 2015). They identified two key challenges. Firstly, retrieving implicit evidence not explicitly mentioned in the query, particularly in cases involving comparisons. For example, statements like "Diarrhoea is the most common adverse event in the primary trial and the secondary trial" require sentences with keywords related to other adverse events than diarrhoea. Secondly, handling negative evidence where no supporting evidence exists, resulting in the retrieval of all sentences. This phenomenon can be attributed to situations where the hypothesis lacks contextual support within the premise. In such cases, presenting the entire premise can provide medical professionals with the necessary justification and rationale for the hypothesis. The authors suggested further research on the necessity of complete evidence in the context of clinical trial data for NLI.

In our research, we explore the significance of contextual representation from the clinical trial section in selecting sentences as evidence for a hypothesis. Context refers to additional linguistic information derived from the source of a candidate sentence, which, in our case, is the CTR section containing the sentence. For instance, Table 1 presents an example where the hypothesis involves a comparison between two cohorts. The highlighted sentences in the table represent the evidence sentences to be retrieved for the hypothesis. The adverse events section of the CTR serves as the premise, and the first sentence, "Adverse Events 1:", is crucial in understanding that the subsequent sentence, "Total: 14/67 (20.90%)," reports the adverse events of cohort 1. Hence, we hypothesize that this relevant information can aid in recognizing the correct evidence. Without fully leveraging contextual information beyond the candidate sentence may potentially lead to sub-optimal solutions in ambiguous cases requiring additional context or references. We study different techniques that can combine the contextual information of the CTR premise, while keeping a single sentence as target evidence, as in related works. Then, we fine-tune pretrained transformer models in various setups and present our findings.In this study, we are the first in examining the role of integrating context into evidence selection for NLI drawn from clinical trials.

Related Work

NLI in CTRs

In the domain of medical research, the evidence retrieval from clinical trials has received significant attention from researchers. Existing endeavors in evidence retrieval from CTRs can be broadly categorized into two main groups. One line of research treats evidence retrieval from CTRs as a classification task, while the other treats it as an information retrieval (IR) task. Chen et al. (2023) considers a classification-oriented approach where the concatenation of the hypothesis and each sentence from a candidate CTR is processed through a pre-trained BioLinkBERT model (Yasunaga, Leskovec, and Liang 2022), which classifies the corresponding sentence as either *evidence* or *not evidence*. Bevan, Turbitt, and Aboshokor (bevan-etal-2023-mdc) employs a similar setup with a more granular approach by considering token level relevance as opposed to the sentence level relevance. Rajamanickam and Rajaraman (2023) use a encoder-decoder based T5 model instead of BERT like encoder models to directly generate the output tokens *relevant* or *irrelevant* from the concatenated hypothesis and candidate sentence.

Conversely, the IR approach treats the task as a retrieval problem, wherein the hypothesis acts as a query and each sentence of the CTR serves as a document. Widely used IR techniques like BM25 (Mahendra, Spina, and Verspoor 2023b) or sentence similarity scores (Jullien et al. 2023b) are employed to retrieve evidence from clinical trials. However, these methods do not incorporate additional contextual information from the CTRs.

In the work by (Zhou et al. 2023), a different perspective is adopted. They consider all sentences of the CTR premise concurrently. Yet, due to the constrained capacity of most transformer models to handle lengthy texts, their approach only incorporates the initial tokens of each sentence. As a result, this approach might overlook crucial information, especially in the context of extensive clinical trials. Differently from previous approaches, our study focus on integrating the context of the CTR premise with the candidate sentence for evidence retrieval.

Context Representation

The integration of context representation has consistently demonstrated enhanced performance across various Natural Language Processing (NLP) tasks. By capturing contextual relationships between words or entities, models can grasp intricate linguistic nuances, leading to improved accuracy and robustness. One of the tasks where contextual modeling has garnered significant attention recently is answer sentence selection for question answering (QA) (Lauriola and Moschitti 2021). This task involves selecting the sentence from a set of candidate answer sentences that correctly responds to a given question. This process aligns with evidence retrieval in NLI, where the goal is to choose the appropriate evidence from a pool of candidate evidences. Furthermore, NLI serves as a foundational task for QA (10). Several studies have demonstrated the benefits of incorporating contextual information in the sentence selection process for QA (Di Liello, Garg, and Moschitti 2023; Van Nguyen et al. 2023). The authors (Lauriola and Moschitti 2021) explored context representation using multiple sentences and Bag-of-Word (BOW) features in Transformer networks, resulting in significant performance improvements for state-of-the-art models. Similarly, the authors (Han, Soldaini, and Moschitti 2021) modeled local context using preceding and succeeding sentences of each answer candidate, and global context using N-gram overlap and cosine similarity and thus improved the answer selection

performance in QA.

Methodology

We initiate our study by introducing the baseline model, following established research practices, which does not encompass contextual factors during its implementation. Subsequently, we elaborate on our approach to embedding context, presenting four distinct strategies: utilization of bag of words (BOW), incorporation of BM25, integration of document summaries, and implementation of bi-directional Long Short-Term Memory (BiLSTM) models.

Baseline

In line with prior work (Jullien et al. 2023b; Vladika and Matthes 2023), we adopted a non-contextualized approach for modeling candidate sequences. Each sequence consisted of a candidate sentence s_i concatenated with the claim c, represented as $a_i = [[CLS]; c; [SEP]; s_i; [SEP]]$. Encoding these sequences using a base language model yielded dense representations $h_i = BERT(a_i)$. We consider T_{CLS} which is the representation of the $[CLS]$ token at the last layer. A Multi-Layer Perceptron (MLP) classifier model is employed to classify the candidate sentence being evidence or not as shown in Fig. 1a.

Context Representation

Bag of Words (BOW) Our approach, inspired by prior work (Lauriola and Moschitti 2021), incorporates context into evidence selection for inference with BOW based features. As mentioned earlier, the CTR premise associated with the candidate sentence, s_i is considered as context for this study. Hence, we extract BOW vector, denoted as T_{BOW} for the CTR section associated with the candidate sentence. T_{BOW} captures the frequency of each input word from the CTR premise. It is to be noted that T_{BOW} has the same number of components as the vocabulary utilized by the base language model. Finally, we normalize the T_{BOW} vector and concatenate it with T_{CLS}, which is the $[CLS]$ representation of the candidate sentence s_i as shown in Fig. 1e. Similar to the baseline model, this representation is now fed to MLP for classification.

BM25 BM25 can retrieve relevance based on exact matching. In the context of information retrieval, studies have demonstrated the effectiveness of globally representative features, like BM25, in capturing the significance of terms within the global context information (Kim et al. 2022). Inspired by previous work (Askari et al. 2023), we inject the BM25 score by modifying the input to the base model as shown in Fig. 1b. The authors extensive experiments suggest that injection of BM25 score in the middle of the input is significantly more effective compared to other approaches. Specifically, each sequence consisted of a candidate sentence s_i concatenated with the claim c, represented as $[CLS]a_i = [c; [SEP]; BM25_{score}; [SEP]; s_i]$. where $BM25_{score}$ is the relevance score between hypothesis and CTR premise. For normalizing BM25 scores, we used min-max standardization.

Summary Summary can be an optimal choice for representation of the context of a document (Lauriola and Moschitti 2021). Hence, we adapt our input sequence as follows: $a_i = [c; [SEP]; s_i; [SEP]; summary_{ctr}]$, where $summary_{ctr}$ denotes the summary of the clinical trial report (CTR) section that encompasses the candidate sentence, generated using a T5 generative model (Raffel et al. 2020). The T5 model has undergone fine-tuning using the *News Summary dataset*[2] with a focus on enhancing its performance for the summarization downstream task. By incorporating the CTR summary into the input sequence, we aim to capture concise yet informative context as shown in Fig 1c.

BiLSTM BiLSTM comprises of two LSTMs: a forward LSTM and a backward LSTM. These components are commonly employed to capture context information in various natural language processing tasks (Tian 2022). Here, we adopt a batch-based processing approach to handle the sentences within the CTR section. Consequently, the batch size dynamically adjusts in each epoch according to the length of the CTR section. As shown in Fig 1e, to capture contextual dependencies within the batch, we introduce a BiLSTM layer. This incorporation of a BiLSTM layer is may enhance the model's ability by capturing contextual dependencies across the batch.

Experimental Setup

Dataset

We conducted experiments using the NLI4CT dataset (Jullien et al. 2023a), which is currently the only available dataset encompassing complete CTRs for NLI. The dataset is divided into training, development, and test subsets. While the training and development subsets are publicly accessible, the absence of ground truth for the test subset leads us to present our findings based on the development subset. We employed an 80:20 split on the provided training set for training and validation. This dataset comprises a compilation of CTRs associated with breast cancer, sourced from *ClinicalTrials.gov*[3], and has undergone annotation by clinical domain experts. The corpus encompasses 2400 statements and CTRs, categorized into four distinct sections: eligibility, intervention, results, and adverse events. The 'eligibility' section outlines the prerequisites participants need to meet to engage in the study trial. In the 'intervention' section, details regarding therapy dosage, frequency, and duration are provided. The 'results' section encompasses participant count, outcome metrics, unit sizes, and findings. Additionally, the 'adverse events' section enumerates the signs and symptoms encountered by patients during the clinical study. In each CTR, there exist one or two patient groups referred to as cohorts or arms. These groups might undergo varying treatments or possess distinct baseline characteristics. The dataset comprises two entry types: "Single" and "Comparison". "Single" entries focus solely on a Primary trial, with evidence contained within its Primary CTR. "Comparison"

[2]https://www.kaggle.com/datasets/sunnysai12345/news-summary

[3]https://clinicaltrials.gov/ct2/

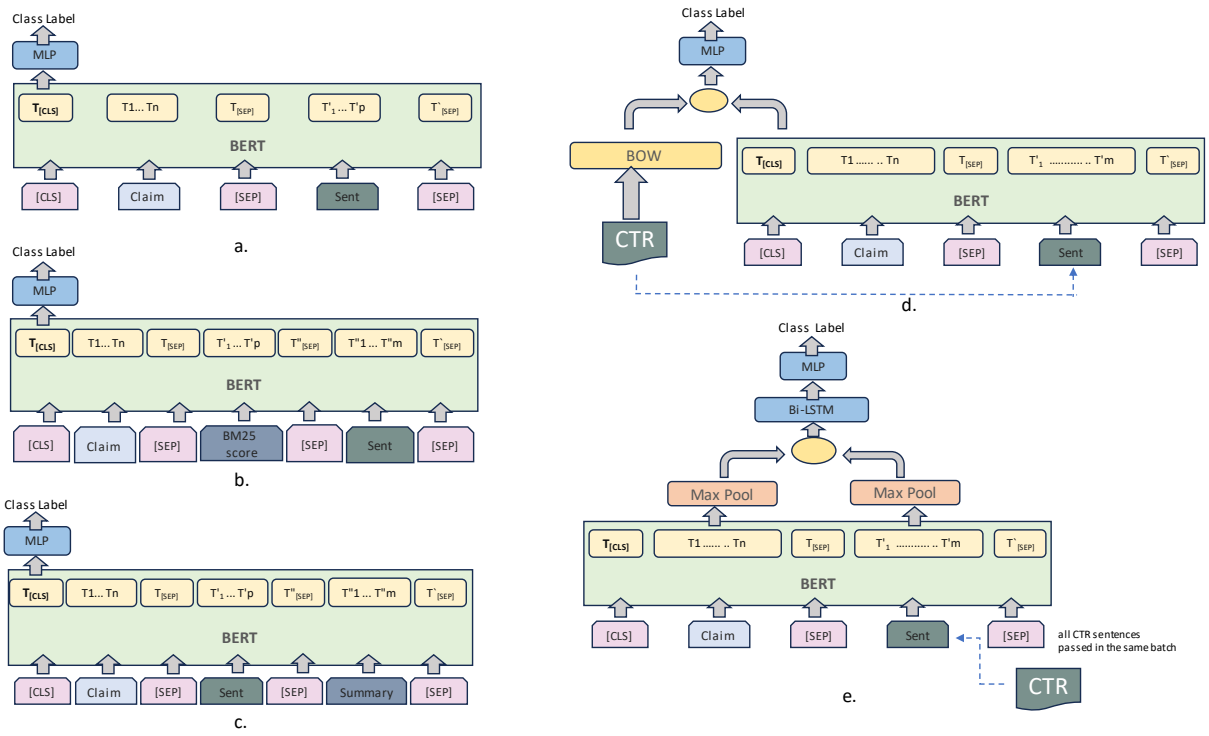

Figure 1: The five approaches evaluated in this work for contextual representation: a. Baseline b. BM25 Score c. Summary d. BOW and e. BiLSTM

Split	Samples	Type		Section				Label	
		Single	Comparison	Intervention	Eligibility	Adverse Events	Results	Evidence	Not Evidence
Train	31948	23675	8273	13594	3412	9809	5133	16713	15235
Dev	7987	5987	2000	3312	828	2577	1270	4151	3836
Test	4224	3388	836	1792	246	1012	1174	1406	2318

Table 2: Description of the Dataset

entries involve both a Primary and a Secondary trial, requiring evidence retrieval from both corresponding CTRs. As discussed earlier we consider each of the sentence as a candidate from the CTR premise and label is binary denoting evidence or not evidence. The dataset details are demonstrated in Table 2

Model Variants

As base language models, we employed DeBERTa-v3 (He, Gao, and Chen 2021) and BioLinkBERT (Yasunaga, Leskovec, and Liang 2022), which have been reported to yield superior performance compared to other language models in prior studies (Jullien et al. 2023b; Vladika and Matthes 2023). The BioLinkBert model is a transformer encoder, akin to BERT, underwent pretraining on PubMed abstracts, alongside citation link information. DeBERTa employs a disentangled attention mechanism, wherein individual words are encoded with two distinct vectors, capturing their content and position. The calculation of attention weights between words is accomplished through disentangled matrices that consider both content and relative positions. In the following sections, the term "BERT" will be used to collectively denote both DeBERTa-v3 and BioLinkBERT.

Implementation Details

All models were downloaded from *HuggingFace*[4]. Input sentences have been tokenized using various tokenizers based on the specific models in use. We fine-tuned thes models using the training dataset with NVIDIA A100 GPUs. The BioLinkBert model and DeBERTa-v3-large model have hidden size of 768 and 1024 respectively. The hyperparameter values were optimized as mentioned in Table 4. The reported

[4]https://huggingface.co/models

Model Name	Method	Accuracy	Precision	Recall	F1
BioLinkBERT	Baseline	0.810	0.764	0.837	0.810
	BOW	0.829	0.820	0.821	0.821
	BM25	0.841	0.822	0.826	0.825
	Summary	0.843	0.790	0.887	**0.836**
	BiLSTM	0.810	0.754	0.856	0.802
DeBERTa-v3-large	Baseline	0.825	0.771	0.869	0.817
	BOW	0.853	0.804	0.890	**0.845**
	BM25	0.834	0.806	0.833	0.820
	Summary	0.844	0.811	0.853	0.832
	BiLSTM	0.831	0.782	0.857	0.818

Table 3: Results on the NLI4CT dataset.

results are the average of three runs using different seeds. For the summarization task using T5, we set the maximum token count to 212. This ensures that when concatenated with the candidate sentence, truncation does not occur. All the codes will be made publicly available upon acceptance of the paper.

Hyperparameters	Value
batch size	16
epoch	8
learning rate	5e-6
weight decay	0.01
warmup ratio	0.06

Table 4: Hyperparameter values used for finetuning the base language models with NLI4CT dataset

Evaluation Metrics

To evaluate, we adopted precision, recall, F1 score, and accuracy, consistent with previous research.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

In the equations TP, TN, FP, FN represent true positive, true negative, false positive and false negative respectively.

Results and Discussion

The results of our experiments on the NLI4CT dataset are presented in Table 3. The integration of contextual information from the CTR premise leads to improved performance in evidence retrieval in most of the cases. For BioLinkBert, the summary-based context representation yields the highest F1 score, while for DeBERTa-v3-large, the BOW-based context representation performs best. The higher recall values in these cases contribute to the overall higher F1 scores. On the other hand, when using the BM25 algorithm, we observe a drop in recall for both context representation approaches. This is expected since BM25 assigns higher scores based on the presence of original terms, penalizing documents that do not contain similar terms. As a result, higher precision values are obtained with BM25. Interestingly, incorporating a BiLSTM layer does not improve the performance compared to the baseline in either context representation case. This is most likely happening due to the significant variation in the length of the CTR premise, leading to batch sizes as small as 2 or as large as 146. This imbalance may pose challenges for the model to effectively learn from the data.

Limitations

There are a few limitations in this study. Firstly, due to the absence of ground truth labels for the original NLI4CT test set, the reported results are based on the development dataset from the original dataset. To address this issue during experimentation, we divided the original training dataset into an 80:20 split for training and development purposes. Furthermore, due to constraints related to GPU and RAM resources, the experiments were carried out with a maximum batch size of 16. This limitation might have restricted our ability to explore the effects of larger batch sizes on the results.

Conclusion

In this study, we have assessed various techniques for introducing contextual elements into evidence retrieval for NLI) from CTRs. Our experimental results clearly demonstrate the superiority of contextual representation over non-contextual approaches. Specifically, we observed that methods based on BOW features and summary-based representations offer the most effective means of integrating context. This research holds the potential to offer medical professionals pertinent evidence sentences without overwhelming them with the entirety of long CTRs, thereby enhancing interpretability. We anticipate that the successful retrieval of evidence will positively impact NLI performance, a promising avenue we are eager to explore in future investigations.

References

Askari, A.; Abolghasemi, A.; Pasi, G.; Kraaij, W.; and Verberne, S. 2023. Injecting the BM25 Score as Text Improves

BERT-Based Re-rankers. In *European Conference on Information Retrieval*, 66–83. Springer.

Bevan, R.; Turbitt, O.; and Aboshokor, M. 2023. MDC at SemEval-2023 Task 7: Fine-tuning Transformers for Textual Entailment Prediction and Evidence Retrieval in Clinical Trials. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1287–1292. Toronto, Canada: Association for Computational Linguistics.

Chen, C.-Y.; Tien, K.-Y.; Cheng, Y.-H.; and Lee, L.-H. 2023. NCUEE-NLP at SemEval-2023 Task 7: Ensemble Biomedical LinkBERT Transformers in Multi-evidence Natural Language Inference for Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 776–781. Toronto, Canada: Association for Computational Linguistics.

DeYoung, J.; Lehman, E.; Nye, B.; Marshall, I. J.; and Wallace, B. C. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Di Liello, L.; Garg, S.; and Moschitti, A. 2023. Context-Aware Transformer Pre-Training for Answer Sentence Selection. *arXiv preprint arXiv:2305.15358*.

Han, R.; Soldaini, L.; and Moschitti, A. 2021. Modeling context in answer sentence selection systems on a latency budget. *arXiv preprint arXiv:2101.12093*.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jullien, M.; Valentino, M.; Frost, H.; O'Regan, P.; Landers, D.; and Freitas, A. 2023a. NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports. *arXiv preprint arXiv:2305.03598*.

Jullien, M.; Valentino, M.; Frost, H.; O'Regan, P.; Landers, D.; and Freitas, A. 2023b. SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data. *arXiv preprint arXiv:2305.02993*.

Kim, M.-Y.; Rabelo, J.; Okeke, K.; and Goebel, R. 2022. Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods. *The Review of Socionetwork Strategies*, 16(1): 157–174.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966. PMLR.

Lauriola, I.; and Moschitti, A. 2021. Answer sentence selection using local and global context in transformer models. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I*, 298–312. Springer.

Mahendra, R.; Spina, D.; and Verspoor, K. 2023a. ITTC at SemEval 2023-Task 7: Document Retrieval and Sentence Similarity for Evidence Retrieval in Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation, Toronto, Canada. Association for Computational Linguistics*.

Mahendra, R.; Spina, D.; and Verspoor, K. 2023b. ITTC at SemEval 2023-Task 7: Document Retrieval and Sentence Similarity for Evidence Retrieval in Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2338–2342. Toronto, Canada: Association for Computational Linguistics.

Mishra, S.; and Sachdeva, B. S. 2020. Do we need to create big datasets to learn a task? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, 169–173.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Rajamanickam, S.; and Rajaraman, K. 2023. I2R at SemEval-2023 Task 7: Explanations-driven Ensemble Approach for Natural Language Inference over Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 1630–1635. Toronto, Canada: Association for Computational Linguistics.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

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

Tian, Y. 2022. Named Entity Recognition in Emergency Domain based on BERT-BILSTM-CRF. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, 817–820. IEEE.

Van Nguyen, M.; KC, K.; Nguyen, T.; Nguyen, T. H.; Chadha, A.; and Vu, T. 2023. Question-Context Alignment and Answer-Context Dependencies for Effective Answer Sentence Selection. *arXiv preprint arXiv:2306.02196*.

Vladika, J.; and Matthes, F. 2023. Sebis at SemEval-2023 Task 7: A Joint System for Natural Language Inference and Evidence Retrieval from Clinical Trial Reports. *arXiv preprint arXiv:2304.13180*.

Yasunaga, M.; Leskovec, J.; and Liang, P. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Zhou, Y.; Jin, Z.; Li, M.; Li, M.; Liu, X.; You, X.; and Wu, J. 2023. THiFLY Research at SemEval-2023 Task 7: A Multi-granularity System for CTR-based Textual Entailment and Evidence Retrieval. *arXiv preprint arXiv:2306.01245*.