KnowComp at SemEval-2023 Task 7: Fine-tuning Pre-trained Language Models for Clinical Trial Entailment Identification

Weiqi Wang*, **Baixuan Xu***, **Tianqing Fang**, **Lirong Zhang**, **Yangqiu Song**
Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
wwangbw@cse.ust.hk, bxuan@connect.ust.hk, tfangaa@cse.ust.hk
lzhangdo@connect.ust.hk, yqsong@cse.ust.hk

Abstract

In this paper, we present our system for the textual entailment identification task as a subtask of the SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data. The entailment identification task aims to determine whether a medical statement affirms a valid entailment given a clinical trial premise or forms a contradiction with it. Since the task is inherently a text classification task, we propose a system that performs binary classification given a statement and its associated clinical trial. Our proposed system leverages a human-defined prompt to aggregate the information contained in the statement, section name, and clinical trials. Pre-trained language models are then finetuned on the prompted input sentences to learn to discriminate the inferential relation between the statement and clinical trial. To validate our system, we conduct extensive experiments with a wide variety of pre-trained language models. Our best system is built on DeBERTa-v3-large, which achieves an F1 score of 0.764 and secures the fifth rank in the official leaderboard. Further analysis indicates that leveraging our designed prompt is effective, and our model suffers from a low recall. Our code and pre-trained models are available at https://github.com/HKUST-KnowComp/NLI4CT.

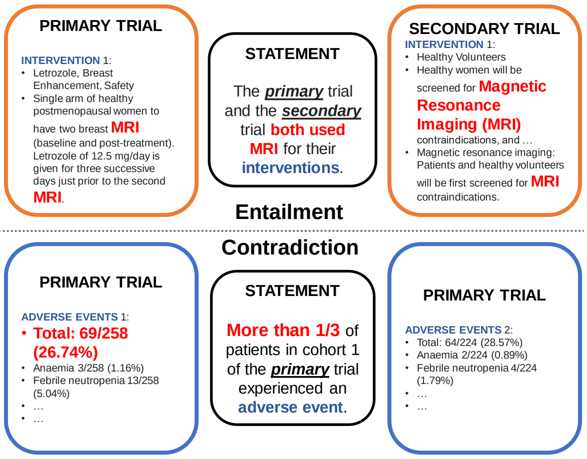
PRIMARY TRIAL
INTERVENTION 1:
• Letrozole, Breast Enhancement, Safety
• Single arm of healthy postmenopausal women to have two breast MRI (baseline and post-treatment). Letrozole of 12.5 mg/day is given for three successive days just prior to the second MRI.
STATEMENT
The primary trial and the secondary trial both used MRI for their interventions.
SECONDARY TRIAL
INTERVENTION 1:
• Healthy Volunteers
• Healthy women will be screened for Magnetic Resonance Imaging (MRI) contraindications, and …
• Magnetic resonance imaging: Patients and healthy volunteers will be first screened for MRI contraindications.
Entailment
Contradiction
PRIMARY TRIAL
ADVERSE EVENTS 1:
• Total: 69/258 (26.74%)
• Anaemia 3/258 (1.16%)
• Febrile neutropenia 13/258 (5.04%)
• …
• …
STATEMENT
More than 1/3 of patients in cohort 1 of the primary trial experienced an adverse event.
PRIMARY TRIAL
ADVERSE EVENTS 2:
• Total: 64/224 (28.57%)
• Anaemia 2/224 (0.89%)
• Febrile neutropenia 4/224 (1.79%)
• …
• …

Figure 1: A demonstration of textual entailment and contradiction between the medical statements and clinical trial records. The statement may claim one or two CTRs on a specific section.

1 Introduction

Recently, the proliferation of Clinical Trial Reports (CTRs) provides a large-scale reference base of scientific and factual knowledge for medical practitioners to make evidence-based clinical diagnoses (Zlabinger et al., 2020). However, it also posits the challenge that clinical practitioners cannot memorize all current literature in order to provide up-to-date personalized evidence-based medical care (DeYoung et al., 2020). With the recent advances in Natural Language Processing (NLP) systems, multiple language models pre-trained in the medical domain have been proposed to tackle medical NLP tasks efficiently (Rasmy et al., 2021; Lewis et al., 2020b; Liu et al., 2022; Kanakarajan et al., 2021). This makes NLP systems more practical and feasible to support the large-scale interpretation and retrieval of medical evidence (Marshall et al., 2020; Molinet et al., 2022; Pradeep et al., 2022; Yasunaga et al., 2022). Though current works study the application of state-of-the-art NLP techniques in the medical domain extensively, evaluation benchmarks are still not comprehensive enough. In this manner, Jullien et al. (2023) propose the Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT) task by constructing an effective evaluation benchmark based on a collection of breast cancer CTRs[1], statements, and explanations. All the collected data are in English, and the labels are annotated by domain experts. Specifically, the annotated statements make some specific types of claims about the informa-

* Equal Contribution

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

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1–9
July 13-14, 2023 ©2023 Association for Computational Linguistics

tion contained in one of the sections in the CTR premise that can be either focusing on one CTR only or comparing a pair of CTRs. The collected benchmark is associated with two proposed subtasks, which address both discrimination and retrieval problems. In this paper, we focus on the first task, textual entailment identification, which aims to determine the inferential relation between a medical statement and the collected clinical trials. Given a statement and CTR pair, the objective is to predict whether the statement affirms a valid entailment of the CTR or forms a contradiction with it, as shown in Figure 1. A system with superior performance on this task can retrieve medical entailment in real-life CTRs and provide clinical practitioners with accurate predictions of treatment outcomes (Katsimpras and Paliouras, 2022), which in turn aids in diagnosis and treatment (Zhang et al., 2020b).

With the recent advancement in Pre-Trained Langauge Models (PTLMs) on text classification tasks (Howard and Ruder, 2018; Wang et al., 2023b), we propose a simple yet effective system that is purely built based on fine-tuning PTLM. By using a carefully designed textual prompt, we aggregate the clinical statement, the claimed section name, and related CTR premises together for the models to learn cohesively. The models are asked to perform binary classification given a prompted input sentence, which stands for discriminating whether the statement claims an entailment or forms a contradiction. To validate the effectiveness of our proposed system, we evaluate a wide variety of PTLMs on both the validation and testing sets and study the ablation of our designed prompt.

Extensive experiment results demonstrate that our system maximally achieves an F1 score of 0.764 and ranks fifth on the official leaderboard. Specifically, DeBERTa-v3 (He et al., 2023b) achieves the best performance and is much more performant than other PTLMs, which highlights its strong language modeling capability. While the large version of other PTLMs may be hard to converge and cannot surpass their respective base version, DeBERTa-v3-large significantly outperforms its smaller version. Further analysis results demonstrate that dropping our designed prompt leads to performance drops, which demonstrates the effectiveness of our designed prompt. However, we observe that our model can only achieve a recall score of 0.772, which ranks thirteenth on the official leaderboard and is a key bottleneck that prevents our system from achieving a higher F1 score. Future works should focus on improving the models' robustness in discriminating true positive examples. Our work studies the performance of various PTLMs on medical multi-evidence natural language reasoning. The experimental results can greatly help clinical practitioners to provide personalized care. We thus make our data, code, and finetuned models publicly available[2] for future contributions.

	Train	Dev	Test
#data	1700	200	500
Avg.Length	19.67	18.68	21.63

Table 1: Statistics of the dataset (Jullien et al., 2023). We report the number of data and the average number of tokens in the statements within each split.

2 Problem Definition

In the textual entailment identification task, each input data contains three components: a medical statement, a section name indicating which section the statement claims about, and one or two CTR records that serve as the evidence to verify the statement. Specifically, if the statement only makes claims about one certain trial, then only the corresponding trial will be used as input data. On the other hand, if the statement claims a comparison between a primary trial and a second trial, both CTRs need to be considered and are provided to the model as input information. Consequently, the textual entailment identification task aims to determine the inferential relation between the medical statement and the associated section(s) in the claimed CTR(s). There are two possible inferential relations for each statement: *entailment* and *contradiction*. Models are expected to predict whether each statement affirms an entailment or forms a contradiction given the associated section from the claimed CTR(s).

Formally, denote the medical statement as t, the section name as s_t, the inferential relation as r_t, and one CTR as $C = \{C_{s1}, C_{s2}, C_{s3}, ..., C_{s|C|}\}$. If two CTRs are involved, they are denoted as C^1 and C^2. The models will be given (t, s_t, C_{s_t}) or $(t, s_t, C^1_{s_t}, C^2_{s_t})$ as input and required to predict y_t that will be compared against r_t.

We use the dataset provided by Jullien et al.

[2] https://github.com/HKUST-KnowComp/NLI4CT

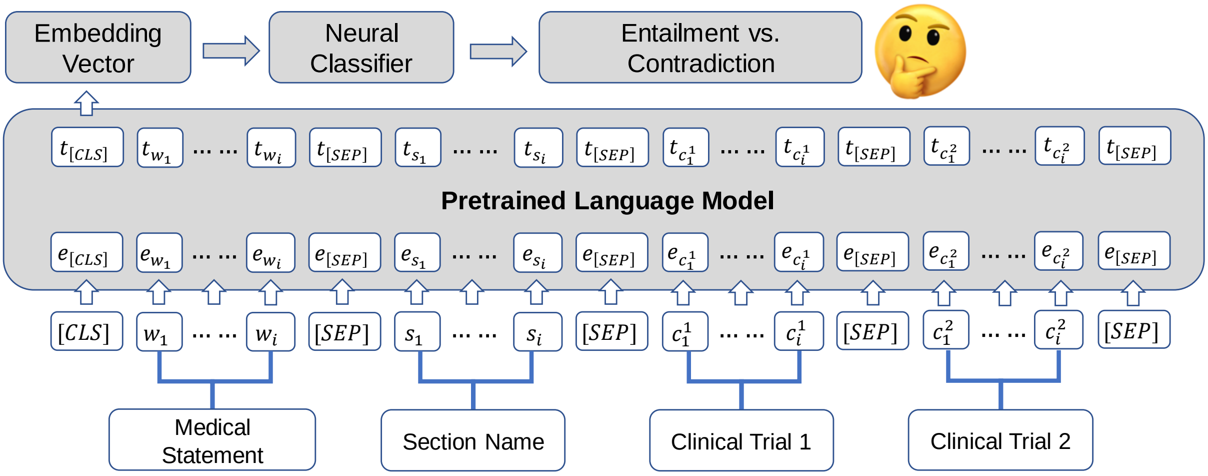
Embedding Vector
Neural Classifier
Entailment vs. Contradiction
$t_{[CLS]}$ t_{w_1} …… t_{w_i} $t_{[SEP]}$ t_{s_1} …… t_{s_i} $t_{[SEP]}$ $t_{c_1^1}$ …… $t_{c_i^1}$ $t_{[SEP]}$ $t_{c_1^2}$ …… $t_{c_i^2}$ $t_{[SEP]}$
Pretrained Language Model
$e_{[CLS]}$ e_{w_1} …… e_{w_i} $e_{[SEP]}$ e_{s_1} …… e_{s_i} $e_{[SEP]}$ $e_{c_1^1}$ …… $e_{c_i^1}$ $e_{[SEP]}$ $e_{c_1^2}$ …… $e_{c_i^2}$ $e_{[SEP]}$
$[CLS]$ w_1 …… w_i $[SEP]$ s_1 …… s_i $[SEP]$ c_1^1 …… c_i^1 $[SEP]$ c_1^2 …… c_i^2 $[SEP]$
Medical Statement
Section Name
Clinical Trial 1
Clinical Trial 2

Figure 2: Overview of our proposed framework with the BERT family as the representative model.

(2023) to study this task. In the dataset, a total of 2,400 statements are split evenly across the different sections and classes. The statements and evidence are generated by clinical domain experts, clinical trial organizers, and research oncologists from the Cancer Research UK Manchester Institute and the Digital Experimental Cancer Medicine Team. Each Clinical Trial Report (CTR) consists of 4 sections: Eligibility criteria, Intervention, Results, and Adverse events. It may contain 1-2 patient groups, which may receive different treatments or have different baseline characteristics. Detailed statistics are provided in Table 1. More explanations regarding the CTRs and sections are provided in Appendix A.

3 System Overview

In this section, we introduce our proposed system. A general sketch is presented in Figure 2.

3.1 Prompt Design

The first step is to aggregate different input components into a whole sentence before performing further classification. Following the definition in Section 2, we extract $t, s_t, C^1_{s_t}, C^2_{s_t}$ and use a discrete prompt to concatenate them together. For each section in a clinical trial, we use commas to link each section's evidence into a paragraph to incorporate the prompt. To efficiently separate different components instead of introducing external natural language guidance, we leverage several predefined separator tokens, denoted as `[SEP]`, to perform the isolation. The exact prompt used for training can then be denoted as "t `[SEP]` s_t `[SEP]` $C^1_{s_t}$ `[SEP]` $C^2_{s_t}$". Other pre-defined special tokens, such as `[CLS]`, `[BOS]`, `[EOS]`, are appended at their appropriate positions. The prompted sentences are further passed as input to the language models.

3.2 Encoder Model Selection

We experiment with a wide collection of popular transformer-based (Vaswani et al., 2017) pretrained language models as text encoders to obtain the embedding of the prompted input sentences.

BERT (Devlin et al., 2019): BERT is the very first bidirectional language model that is purely based on transformer architecture. It is pre-trained by using Masked Language Modeling (MLM) and Next-Sentence Prediction (NSP) objectives.

BioClinical-BERT (Alsentzer et al., 2019): BioClinical-BERT is a BERT model that is pretrained using the medical notes from a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston. It is considered to be a domain-specific language model that is potentially effective in clinical NLP tasks.

ALBERT (Lan et al., 2020): ALBERT is a variant of BERT that leverages two parameter reduction techniques, factorized embedding parameterization and cross-layer parameter sharing, to improve its parameter-efficiency. A sentence order prediction loss during pretraining is also introduced to replace the original NSP objective.

BART (Lewis et al., 2020a): BART is an encoder-decoder model that uses a bidirectional encoder and an autoregressive decoder to perform sequence-to-sequence modeling. It is pre-trained with the objectives of corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. It has been shown to

Backbone PTLM / Method	Validation			Testing		
	F1	Prec.	Rec.	F1	Prec.	Rec.
BERT-base *110M*	69.2	60.4	81.0	58.1	58.2	58.0
BERT-large *340M*	70.9	61.3	84.0	61.5	57.5	66.0
BioClinical-BERT-base *110M*	65.3	56.1	78.0	59.6	58.5	60.8
ALBERT-v2-base *12M*	67.1	50.5	100.0	64.4	55.3	77.2
ALBERT-v2-large *18M*	67.1	50.8	99.0	66.7	50.0	100.0
ALBERT-v2-xlarge *60M*	67.1	50.8	99.0	66.8	50.1	100.0
ALBERT-v2-xxlarge *235M*	67.1	50.5	100.0	66.2	50.1	97.6
BART-base *139M*	67.3	50.8	100.0	65.6	49.9	95.6
BART-large *406M*	66.9	50.7	98.0	63.8	58.5	70.0
RoBERTa-base *110M*	70.7	62.8	81.0	60.7	58.0	63.6
RoBERTa-large *340M*	67.6	63.7	72.0	56.5	59.2	54.0
DeBERTa-v3-base *214M*	75.8	86.0	67.7	65.6	65.2	66.0
DeBERTa-v3-large *435M*	**81.5**	75.9	88.0	**76.4**	75.7	77.2
ELECTRA-base *110M*	70.3	78.0	63.9	61.4	60.5	62.4
ELECTRA-large *340M*	76.1	71.7	81.0	66.5	70.7	62.8
GPT2-base *117M*	39.0	31.0	52.5	60.3	51.4	72.8
GPT2-medium *345M*	44.2	38.0	52.8	64.6	50.8	88.8
GPT2-large *774M*	61.5	60.0	63.2	56.0	55.5	56.4

Table 2: Full experiment results (%) of the textual entailment identification task on both validation and testing sets by various language models. We report the F1, precision (Prec.), and recall (Rec.) scores for every model. The best performances with respect to F1 scores are bold-faced.

be effective on both generative and discriminative NLP tasks.

RoBERTa (Liu et al., 2019): RoBERTa is a variant of BERT that is obtained by using an improved recipe for training BERT models. The training takes advantage of dynamic masking, full sentences without NSP loss, large mini-batches, and a large byte-level byte-pair encoding strategy. As a result, the performance of RoBERTa is significantly improved compared with the traditional BERT model.

ELECTRA (Clark et al., 2020): ELECTRA introduces a new way of pretraining MLM language models by corrupting the input by replacing some tokens with plausible alternatives sampled from a small generator network instead of performing the traditional input masking. In addition, the model is asked to predict whether each token in the corrupted input was replaced by a generator sample instead of predicting the original identity of the masked tokens. It is much more efficient than previous pretraining approaches and can achieve comparable results.

DeBERTa (He et al., 2021, 2023b): DeBERTa is the current state-of-the-art language model that improves the BERT and its variants by using disentangled attention and enhanced mask decoder. Meanwhile, the training efficiency of DeBERTa is also improved using ELECTRA-Style pre-training with gradient disentangled embedding sharing. Performances on downstream natural language understanding tasks are significantly improved compared with previously introduced models.

GPT2 (Radford et al., 2019; Brown et al., 2020): GPT2 is a generative language model that is pretrained using a Causal Language Modeling (CLM) objective. In addition, the model takes a batch of sequences of the continuous text of a certain length as inputs, and the respective targets are the same sequence with one token shifted to the right. It also uses a masking mechanism to ensure that the prediction of a specific token is only based on previous existing tokens instead of future tokens. Such autoregressive training enables GPT2 to be extremely powerful in generating texts.

Considering the unique design and training objective function of each model, we adopt different extraction strategies for different models to extract the representation embedding of the input sentence from the token embeddings encoded by the language models. Specifically, for language models with a masked language modeling objective, including BERT, ALBERT, RoBERTa, DeBERTa, and ELECTRA, the embedding of the `[CLS]` token is used as the representation vector. While for autoregressive generative models such as BART and GPT2, the embedding of the last token in the decoder is used as the representation of the input sequence. Such a strategy best utilizes the information provided by the encoder PTLM (Fang et al., 2022).

Two linear layers are connected after the encoder language model to perform binary classification on the representation vectors. Tanh function is used as the activation function (Nwankpa et al., 2018) and two dropout (Srivastava et al., 2014) layers are added appropriately to avoid overfitting.

3.3 Model Training

Denote the prompted input sentence as x_i with $|x_i|$ tokens, the inferential relation label as y_i. All models, denoted as θ, are trained using a standard cross-entropy loss, as shown in Equation 1. The predictions of x_i, denoted as $\theta(x_i)$, will be used to compute the loss against the truth label y_i.

$$L(x_i, \theta) = -\sum_{i=1}^{|x|} y_i \log(\theta(x_i)) \qquad (1)$$

4 Experimental setup

The data split for training, validation, and testing sets follows the original split as released by Jullien et al. (2023). Our system is built upon the Huggingface Transformers[3] Library (Wolf et al., 2020). pre-trained tokenizers and language models are applied directly for further finetuning. The training and evaluation codes are mainly adapted from Fang et al. (2021b,a, 2023). We use a default learning rate of 5e-6 with a batch size of 4 to train the models. An AdamW (Loshchilov and Hutter, 2019) optimizer is used to update the parameters. The max sequence length for the tokenizer is set to 512, which is the most common longest input length for PTLMs. The models are evaluated on the validation every 10 steps by using precision, recall, and F1 scores (Powers, 2020). Early stopping is used where the best checkpoint is selected when the largest validation F1 score is achieved. The best checkpoint is further submitted to the CodaLab platform to acquire test set performances. We train our models with a sufficiently large number of epochs to ensure that underfitting does not occur. All experiments are repeated three times using different random seeds, and the average performance is reported. Four NVIDIA RTXA6000 (48G) and four NVIDIA RTX3090 (24G) graphical cards are used as the computational infrastructures. The number of parameters for every model is reported in Table 2. All testing set performances are acquired through submissions on the official CodaLab platform.

5 Results

The full results are shown in Table 2. We can observe that PTLMs with a Masked Language Modeling (MLM) objective generally can achieve satisfactory performance. The majority of them achieve an F1 score of above 0.6 on the test set. The GPT2 family, on the other hand, struggles with the task and can only achieve comparable performances when the parameter is over 700 million, which is nearly seven times more than other discriminative PTLMs. One possible reason is that autoregressive language modeling is not that competitive at classification tasks, and the GPT2 model cannot learn the negative samples well. Our best model is finetuned based on the DeBERTa-v3-large, which reaches a 0.764 F1 score and is significantly outperforming other models. This may be due to the fact that DeBERTa-v3-large possesses the largest number of parameters, and its disentangled parameter-sharing technique is effective for evidence-understanding tasks (He et al., 2023b). Such a result enables our model to secure a fifth rank on the official leaderboard and implies that advanced PTLM can be proficient in solving the textual entailment identification task.

PTLM	F1.	Prec.	Rec.
DeBERTa-v3-large	**76.4**	**75.7**	**77.2**
⋄ w/o **Special Token**	73.4	73.5	73.2
⋄ w/o **Statement**	61.1	57.1	65.6

Table 3: Ablation study on our prompt used for information aggregation. w/o stands for dropping a specific component. Prec. and Rec. refer to precision and recall, respectively.

We further study the ablation of our proposed prompt. DeBERTa-v3-large finetuned on two different input prompts are compared as baselines. **Special Token** stands for replacing the separator tokens `[SEP]` with natural language guidance such as "*the evidence from CTR1 is*". **Statement** stands for using the statement only for prediction. The results, shown in Table 3, support our claim that leveraging predefined special tokens as separators and concatenating evidence after the statement is

[3] https://huggingface.co/docs/transformers

effective.

Meanwhile, our best model's precision score is ranked fourth, and the recall score is ranked thirteenth. This indicates that the model is wrongly classifying many entailments as contradictions, which causes a relatively low recall and is considered a major limitation of our system. By further observing the errors made by our system, we find that challenging CTR-statement pairs involving complex or mathematical reasoning (Ferreira and Freitas, 2020) are often wrongly classified. In addition, due to the length of some CTR premises being much longer than the PTLM's input max length, important pieces of evidence may be truncated and failed to provide guidance when making inferences on the statement. In our paper, we ignore this issue due to the fact that such cases infrequently occur through our manual inspection. However, positional encoding techniques such as relative position representation (Shaw et al., 2018) should be implemented to compensate for such information loss.

In addition, we foresee three lines of future works that can be addressed to enhance the system's robustness further: (1) Leveraging evidence retrieval (Yadav et al., 2021; Conforti et al., 2020) or abstractive summarization (de Vargas Feijó and Moreira, 2023; Mao et al., 2022) to reduce the noise caused by irrelevant evidence and present the model with the most influential evidence. (2) Applying augmentation techniques (He et al., 2022; Wang et al., 2023a; Wu et al., 2022; Qin and Joty, 2022) and aligning medical inferences against large eventuality knowledge graphs (Zhang et al., 2022, 2020a; Röder et al., 2018) to mine inferential knowledge scalably and train the model more robustly. (3) Leveraging GPT3.5 (Brown et al., 2020; Ouyang et al., 2022), ChatGPT[4] or other advanced large language models to generate or extract (He et al., 2023a) key evidence as they have shown strong capabilities in these tasks (Chan et al., 2023).

6 Conclusion

In this paper, we propose to finetune pre-trained language models to tackle the task of textual entailment identification as a solution to SemEval2023 Task 7. By using a discrete prompt to aggregate the medical statement and CTR premises, we maximally achieve an F1 score of 0.764 and ranked fifth in the official leaderboard. Analysis experiments also prove the effectiveness of our proposed prompt. However, our system suffers from an excessively low recall rate. This indicates that the model misclassifies too many entailments as contradictions. Future work can focus on improving the ability of the model to identify entailments and improve its robustness. Our work demonstrates that applying a larger pre-trained language model can effectively identify clinical trial entailments, which can be applied to real-world scenarios. All of our code, data, and models are publicly available for future contributions.

[4] https://chat.openai.com/chat

Acknowledgements

The authors would like to thank the committee of SemEval2023, the organizers of Task 7, and the anonymous reviewers. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong, the MHKJFS (MHP/001/19) from ITC of Hong Kong and the National Key R&D Program of China (2019YFE0198200) with special thanks to HKMAAC and CUSBLT, and the Jiangsu Province Science and Technology Collaboration Fund (BZ2021065). We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. STANDER: an expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4086–4101. Association for Computational Linguistics.

Diego de Vargas Feijó and Viviane P. Moreira. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artif. Intell. Law*, 31(1):91–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin E. Nye, Iain James Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020*, pages 123–132. Association for Computational Linguistics.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. Ckbp v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Quyet V. Do, Hongming Zhang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. Pseudoreasoner: Leveraging pseudo labels for commonsense knowledge base population. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3379–3394. Association for Computational Linguistics.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Deborah Ferreira and André Freitas. 2020. Premise selection in natural language mathematical texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7365–7374. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023a. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023b. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *11th International Conference on Learning Representations, ICLR 2023*. OpenReview.net.

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

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical*

Language Processing, BioNLP@NAACL-HLT 2021, Online, June 11, 2021, pages 143–154. Association for Computational Linguistics.

Georgios Katsimpras and Georgios Paliouras. 2022. Predicting intervention approval in clinical trials through multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1947–1957. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop, ClinicalNLP@EMNLP 2020, Online, November 19, 2020*, pages 146–157. Association for Computational Linguistics.

Lang Liu, Junxiang Ren, Yuejiao Wu, Ruilin Song, Zhen Cheng, and Sibo Wang. 2022. A pre-trained language model for medical question answering based on domain adaption. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part II*, volume 13552 of *Lecture Notes in Computer Science*, pages 216–227. Springer.

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*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Hassan Awadallah, and Dragomir R. Radev. 2022. DYLE: dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1687–1698. Association for Computational Linguistics.

Iain James Marshall, Benjamin E. Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C. Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *J. Am. Medical Informatics Assoc.*, 27(12):1903–1912.

Benjamin Molinet, Santiago Marro, Elena Cabrio, Serena Villata, and Tobias Mayer. 2022. ACTA 2.0: A modular architecture for multi-layer argumentative analysis of clinical trials. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5940–5943. ijcai.org.

Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. 2018. Activation functions: Comparison of trends in practice and research for deep learning. *CoRR*, abs/1811.03378.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

David M. W. Powers. 2020. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061.

Ronak Pradeep, Yilin Li, Yuetong Wang, and Jimmy Lin. 2022. Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2325–2330. ACM.

Chengwei Qin and Shafiq R. Joty. 2022. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2776–2789. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Medicine*, 4.

Michael Röder, Giorgos Stoilos, David Geleta, Jetendr Shamdasani, and Mohammad Khodadadi. 2018. Medical knowledge graph construction by aligning large biomedical datasets. In *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018*, volume 2288 of *CEUR Workshop Proceedings*, pages 218–219. CEUR-WS.org.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. *CoRR*, abs/2305.14869.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. *CoRR*, abs/2305.04808.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Meng Lin, Liangjun Zang, and Songlin Hu. 2022. Text smoothing: Enhance various data augmentation methods on text classification tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 871–875. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2021. If you want to go far go together: Unsupervised joint candidate evidence retrieval for multi-hop question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4571–4581. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8003–8016. Association for Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artif. Intell.*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020a. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Xingyao Zhang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2020b. Patient-trial matching with deep embedding and entailment prediction. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1029–1037. ACM / IW3C2.

Markus Zlabinger, Marta Sabou, Sebastian Hofstätter, and Allan Hanbury. 2020. Effective crowd-annotation of participants, interventions, and outcomes in the text of clinical trial reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3064–3074. Association for Computational Linguistics.

A Dataset Description

In the dataset, the CTRs are collected from a database of privately and publicly funded clinical studies conducted around the world. The statements mainly claim on four sections of these CTRs: **Eligibility Criteria**: A set of conditions for patients to be allowed to take part in the clinical trial. **Intervention**: Information concerning the type, dosage, frequency, and duration of treatments being studied. **Results**: Number of participants in the trial, outcome measures, units, and the results. **Adverse Events**: These are signs and symptoms observed in patients during the clinical trial.