CLaC at SemEval-2024 Task 2: Faithful Clinical Trial Inference

Jennifer Marks[1] **MohammadReza Davari**[1,2] **Leila Kosseim**[1]

[1] Concordia University [2] Mila – Quebec AI Institute

jennifer.marks@mail.concordia.ca
{mohammadreza.davari,leila.kosseim}@concordia.ca

Abstract

This paper presents the methodology used for our participation in SemEval 2024 Task 2 (Jullien et al., 2024) – *Safe Biomedical Natural Language Inference for Clinical Trials*. The task involved Natural Language Inference (NLI) on clinical trial data, where statements were provided regarding information within Clinical Trial Reports (CTRs). These statements could pertain to a single CTR or compare two CTRs, requiring the identification of the inference relation (entailment vs contradiction) between CTR-statement pairs. Evaluation was based on F1, Faithfulness, and Consistency metrics, with priority given to the latter two by the organizers. Our approach aims to maximize Faithfulness and Consistency, guided by intuitive definitions provided by the organizers, without detailed metric calculations. Experimentally, our approach yielded models achieving maximal Faithfulness (top rank) and average Consistency (mid rank) at the expense of F1 (low rank). Future work will focus on refining our approach to achieve a balance among all three metrics.

1 Introduction

Clinical trials serve as the cornerstone for evaluating the efficacy and safety of novel medical interventions, playing a pivotal role in advancing healthcare practices (Avis et al., 2006). Clinical Trial Reports (CTRs) encapsulate crucial information regarding trial methodologies and outcomes, serving as indispensable resources for healthcare professionals in treatment decision-making (Bastian et al., 2010). However, the sheer volume of available CTRs, coupled with their rapid proliferation, poses significant challenges for comprehensive literature review and evidence synthesis in clinical practice (DeYoung et al., 2020). Natural Language Inference (NLI) emerges as a promising approach to address this issue (Bowman et al., 2015; Devlin et al., 2018; Raffel et al., 2020), facilitating the scalable interpretation and retrieval of medical evidence (Davari et al., 2020; Sutton et al., 2020; Davari et al., 2019). The SemEval 2024 Task 2 (Jullien et al., 2024) on *Safe Biomedical NLI for Clinical Trials* extends this paradigm to enable automated inference of relationships between statements and CTRs, thus streamlining evidence extraction and enhancing decision-making processes in healthcare.

The 2024 task is a continuation of the one introduced by Jullien et al. (2023b,a), specifically it focuses on Track 1, which focuses on NLI in the context of clinical trials. In this task, the input consists of pairs of Clinical Trial Reports (CTRs) and corresponding statements, where the statements make claims about the information contained within the CTRs. The objective is to determine the inference relation between each CTR-statement pair, classifying them as either entailing or contradicting each other. For instance, given a statement "Drug X is effective in treating condition Y" and a CTR outlining a clinical trial testing Drug X's efficacy, the task is to determine whether the statement is entailed by the CTR or contradicted by it. The datasets used are similar to those introduced by Jullien et al. (2023b), and further details can be found in there work.

Our system primarily focuses on maximizing Faithfulness and Consistency in the context of SemEval 2024 Task 2 (Jullien et al., 2024). To achieve this goal, we adopt a strategy centered around introducing controlled input noise during model training. This approach is based on the hypothesis that a certain level of tolerance towards input perturbations could enhance the faithfulness and consistency of the generated models. Specifically, we experiment with randomly masking a percentage ($k\%$) of tokens in both Clinical Trial Reports (CTRs) and the corresponding statements, thereby exposing the model to varying degrees of input uncertainty.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1673–1677
June 20-21, 2024 ©2024 Association for Computational Linguistics

Through these experiments, we aim to optimize model performance in capturing the relationships between statements and CTRs, ultimately improving the system's effectiveness in clinical trial inference tasks.

Through our participation in SemEval 2024 Task 2 (Jullien et al., 2024), we observed a notable trade-off between different evaluation metrics. While our approach successfully improved Faithfulness and Consistency metrics, it came at the expense of F1 scores. This finding underscores the challenge of balancing these evaluation criteria and thus the need for future refinement to achieve a more harmonious optimization across all relevant metrics. Specifically, our models achieved top-ranking levels of Faithfulness but demonstrated only average performance in Consistency metrics, resulting in lower ranks in F1 assessment. See Sec. 4 for details.

2 System Overview

In our system, we leverage BART (Lewis et al., 2019) as the primary model for all experiments due to its robustness and effectiveness in various natural language processing tasks, particularly in Natural Language Inference (NLI) (Lewis et al., 2019; Barker et al., 2021; Farahnak et al., 2020). To streamline the fine-tuning process and enhance efficiency, we adopt the LoRA technique proposed by Hu et al. (2021), which significantly reduces the fine-tuning time without sacrificing performance. Additionally, we incorporate the Contrastive Tension loss function introduced by Carlsson et al. (2020). to guide the fine-tuning process. This loss function promotes contrastive learning by separately encoding the Clinical Trial Reports (CTRs) and their associated statements using two copies of BART (Lewis et al., 2019) during each training instance. By allowing only one copy to update its parameters at a time, the model is encouraged to focus on learning the essential semantic relationships between the CTRs and the statements.

Moreover, we introduce a novel approach to enhance the robustness of the model by incorporating random token masking during each training instance. Specifically, we randomly mask a percentage ($k\%$) of tokens in both the CTRs and their associated statements. This introduces noise in the input data, forcing the model to adapt to varying degrees of input uncertainty and preventing it from relying solely on superficial patterns. The rationale behind this approach is to encourage the model to concentrate on the fundamental semantic content of the input rather than exploiting surface-level correlations.

Balancing between the three required metrics—Faithfulness, Consistency, and F1—proved to be the primary challenge in our experimental setup. While optimizing for one metric often led to improvements in its performance, it frequently came at the expense of others. We explored different strategies to strike a balance between these metrics. However, finding an optimal solution that simultaneously maximized all three metrics remained unsolved. Our system struggled to maintain high levels of F1 score while simultaneously improving Faithfulness and Consistency metrics. Further exploration of optimization strategies and leveraging ensemble methods may offer potential avenues for achieving a better balance between the metrics.

3 Experimental setup

Training Details The training, validation, and test data for our experiments were all provided by the organizers of the SemEval 2024 Task 2, as outlined by Jullien et al. (2024). We trained our model for a total of 40 epochs, employing the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001, with a batch size of 32. To stabilize and accelerate training, we implemented gradient clipping (Zhang et al., 2019) with a maximum norm of 1.

Additionally, we incorporated a linear warmup stage consisting of 40 gradient steps followed by a Cosine Annealing learning rate schedule (Loshchilov and Hutter, 2016). This strategy enabled gradual adjustment of the learning rate during the initial phase of training, allowing the model to converge more smoothly towards an optimal solution.

Furthermore, we limited the maximum sequence length to 256 tokens for both CTRs and their corresponding statements, aligning with the model architecture and computational capabilities. Consequently, each sequence pair was truncated to a total maximum of 512 tokens. This limitation on sequence length helped in handling memory constraints and optimizing the processing of input sequences during training. Exploring larger context sizes could be advantageous for future improvements in the task.

Evaluation Metrics Our system's performance is measured through three key metrics: (1) Faithfulness, (2) Consistency, and (3) F1 score.

Faithfulness: Faithfulness measures the degree to which a given system arrives at the correct prediction for the correct reason. Intuitively, this is estimated by assessing the model's ability to correctly change its predictions when subjected to a semantic altering intervention. Let N denote the number of statements x_i in the contrast set (C), y_i represent their respective original statements, and $f()$ denote the model predictions. Faithfulness is computed using Equation below:

$$\text{Faithfulness} = \frac{1}{N} \sum_{1}^{N} |f(y_i) - f(x_i)| \quad (1)$$

where $x_i \in C$, $\text{Label}(x_i) \neq \text{Label}(y_i)$, and $f(y_i) = \text{Label}(y_i)$.

Consistency: Consistency aims to measure the extent to which a given system produces the same outputs for semantically equivalent problems. It assesses the system's ability to predict the same label for original statements and contrast statements for semantic preserving interventions. Even if the final prediction is incorrect, the representation of the semantic phenomena should be consistent across the statements. Let N denote the number of statements x_i in the contrast set (C), y_i represent their respective original statements, and $f()$ denote the model predictions. Consistency is computed using Equation below:

$$\text{Consistency} = \frac{1}{N} \sum_{1}^{N} 1 - |f(y_i) - f(x_i)| \quad (2)$$

where $x_i \in C$, $\text{Label}(x_i) = \text{Label}(y_i)$.

F1: The F1 score is a commonly used metric in NLP tasks (Yang et al., 2023; Davari, 2020), measuring the balance between precision and recall of a model's predictions. It is calculated based on the geometric mean of precision and recall, where precision represents the ratio of true positive predictions to the total number of predicted positive instances, and recall represents the ratio of true positive predictions to the total number of actual positive instances.

4 Results

Our experimental results demonstrate a clear trade-off between the level of token masking (k) and the performance metrics of F1, Faithfulness, and Consistency. As we increase the masking level, we observe a consistent trend of decreasing F1 scores alongside increasing Faithfulness and Consistency metrics.

For $k = 0$, representing no token masking, we observe an F1 score of 0.65 (ranking 22 out of 32), a Faithfulness score of 0.51 (ranking 21 out of 28), and a Consistency score of 0.54 (ranking 25 out of 30). As we progressively increase k, we observe a gradual change in the metrics, specifically decrease in F1, and increase of the other 2 metrics. At $k = 30\%$, the highest masking level tested, we observe a significant drop in F1 score to 0.06 (ranking 28 out of 31), accompanied by substantial increases in Faithfulness (0.95, ranking 1 out of 28) and Consistency (0.6, ranking 22 out of 32) metrics.

Given the substantial decrease in F1 scores beyond a masking level of $k = 30\%$, we did not explore higher masking levels. This decision was made due to the observed trade-off, where increasing token masking beyond a certain threshold led to disproportionately low F1 scores, potentially indicating a loss of model generalization and predictive performance.

5 Conclusion

Our experiments underscore the intricate balance between token masking levels and performance metrics in biomedical NLI for clinical trials. We observed a discernible trade-off: while increasing token masking improves Faithfulness and Consistency, it results in diminished F1 scores. This finding highlights the necessity of exploring future approaches that could better optimize the model with multiple evaluation criteria as their objective function. Additionally, another potential avenue for improvement would involve examining alternative metrics to provide further insights into the behaviour of the model (Davari et al., 2022a; Farahnak et al., 2021; Steck et al., 2024; Davari et al., 2022b). Based on our findings, one avenue of research involves refining token masking strategies to achieve a more optimal balance between F1, Faithfulness, and Consistency metrics. Furthermore, exploring ensemble methods and alternative fine-tuning strategies could provide valuable insights into enhancing the overall performance of the model.

Acknowledgements

The authors would like to thank the organisers of the SemEval shared task and the anonymous reviewers for their comments on the previous version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Nancy E Avis, Kevin W Smith, Carol L Link, Gabriel N Hortobagyi, and Edgardo Rivera. 2006. Factors associated with participation in breast cancer treatment clinical trials. *Journal of Clinical Oncology*, 24(12):1860–1867.

Ken Barker, Parul Awasthy, Jian Ni, and Radu Florian. 2021. Ibm mnlp ie at case 2021 task 2: Nli reranking for zero-shot text classification. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 193–202.

Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International conference on learning representations*.

MohammadReza Davari. 2020. *Neural Network Approaches to Medical Toponym Recognition*. Ph.D. thesis, Concordia University.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. 2022a. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16712–16721.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2022b. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*.

MohammadReza Davari, Leila Kosseim, and Tien Bui. 2020. Timbert: toponym identifier for the medical domain based on bert. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 662–668.

MohammadReza Davari, Leila Kosseim, and Tien D Bui. 2019. Toponym identification in epidemiology articles–a deep learning approach. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 26–37. Springer.

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

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*.

Farhood Farahnak, Elham Mohammadi, MohammadReza Davari, and Leila Kosseim. 2021. Semantic similarity matching using contextualized representations. In *Canadian Conference on AI*, volume 1.

Farhood Farahnak, Laya Rafiee, Leila Kosseim, and Thomas Fevens. 2020. Surface realization using pretrained language models. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 57–63.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023a. Nli4ct: Multi-evidence natural language inference for clinical trial reports. *arXiv preprint arXiv:2305.03598*.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? *arXiv preprint arXiv:2403.05440*.

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

Zachary Yang, Yasmine Maricar, Mohammadreza Davari, Nicolas Grenon-Godbout, and Reihaneh Rabbany. 2023. Toxbuster: In-game chat toxicity buster with bert. *arXiv preprint arXiv:2305.12542*.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2019. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*.