CSE517: Project Proposal
Team Cookies: Mingcheng Wang, Eric Xia, Trevor Jung

Citation and Hypotheses:

We wish to verify the Decision-Focused Summarization proposed in this paper [1], which is used to summarize relevant information for a decision, outperforming other text-only summarization methods and model-based explanation methods in decision faithfulness and representativeness.

Usage of Data from Paper and Existing Code:
We will use the yelp dataset described in the paper, freely available online at https://www.yelp.com/dataset/download. Code provided in the paper will also be used to aid in our reproduction, available at https://github.com/ChicagoHAI/decsum.

Discussion of feasibility:
We will attempt to mirror the methology described in the paper as follows. Given an input text $X = \{x_s\}_{s=1}^{s=S}$, where S is the number of sentences, we wish to select a subset of sentences $\tilde{X} \subset X$ to support making a decision y. Thus, the training set, $D_{train} = \{(X_i, y_i)\}$, will be processed in such a way as to find defining traits in a text that lead to a certain decision y. The yelp dataset is defined as follows: for each restuarant, define X as the text body of the first k reviews and let y be the average rating of the first t reviews, where $t > k$ s.t. the problem is to predict future ratings. In the paper, the research group used $k = 10, t = 50$ s.t. their task was to find sentences from a given restaurant's first 10 reviews that supported a prediction of its future rating after 50 reviews.

At this level of abstraction of our project to reproduce the paper, feasibility seems high, given that the yelp data they used is publicly available online, and their documentation for their code is on Github and seems to be well-documented. Possible areas to look into beyond the paper at this level of abstraction include choosing different amounts of data from the yelp dataset and different t and k values to see their effect on the model's performance; this falls under adding new ablations and hyperparameter tuning.

Taking a step down the abstraction ladder, we see that the authors of the paper used a regression model, fine-tuned with *Longformer*. Finally, *DecSum*'s performance on the regression model was compared against certain other text-only summarization methods and model-based explanations.

As mentioned on the authors' Github page for this paper, training the *Longformer* model takes about three hours with half-precision, the step running *DecSum* taking about 10 hours, and the final step getting decision scores for individual sentences based on *DecSum* taking about an hour, all run on a single RTX3090 GPU. Thus, this experiment might have required 14 hours to run on the RTX3090. At this lower level of abstraction, feasibility of methodology still seems mostly certain, but the feasibility of computation seems less certain, as we do not know how our combined $150 of Google Cloud credits and 45G of GPU usage with the CSE machines translates to actual runtime of the model with GPUs.

One exception is that in our reproduction we will not be able to implement the human evaluation used in the paper; this evaluation used Amazon Mechanical Turk to solicit human guesses on which restaurants would be rated higher after participants were shown 50 reviews.

References

[1] Chao-Chun Hsu and Chenhao Tan. Decision-focused summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 117–132, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.